Annette Klosa-Kückelhaus, Stefan Engelberg,
Christine Möhrs, Petra Storjohann (eds.)

Dictionaries and Society

EURALEX XX 22

Proceedings of the
XX EURALEX International Congress,
12-16 July 2022,
Mannheim, Germany

IDS IDS-Verlag

Bibliografische Information der Deutschen Nationalbibliothek

Die Deutsche Nationalbibliothek verzeichnet diese Publikation in der Deutschen Nationalbibliografie; detaillierte bibliografische Daten sind im Internet über http://dnb.dnb.de abrufbar.

IDS-Verlag

LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

Leibniz Gemeinschaft

IDS–Verlag · Leibniz–Institut für Deutsche Sprache · R 5, 6–13 · 68161 Mannheim

Redaktion: Melanie Kraus
Satz: Annett Patzschewitz, Joachim Hohwieler
Titelbild: Norbert Cußler-Volz

Dieses Werk ist unter der Creative-Commons-Lizenz 3.0 (CC BY–SA 3.0) veröffentlicht.
Die Umschlaggestaltung unterliegt der Creative-Commons-Lizenz CC BY-ND 3.0.

Die Online-Version dieser Publikation ist auf den Webseiten des Leibniz–Instituts für Deutsche Sprache (www.ids-mannheim.de) dauerhaft frei verfügbar (Open Access). doi: https://doi.org/10.14618/phpy-6r66

Die gesetzliche Verpflichtung über die Ablieferung digitaler Publikationen als Pflichtexemplare wird durch die manuelle Ablieferung der Netzpublikation an die Badische Landesbibliothek (BLB) erfüllt.

ISBN: 978-3-937241-87-6 (PDF)

© 2022 Annette Klosa-Kückelhaus/Stefan Engelberg/Christine Möhrs/Petra Storjohann

FOREWORD

The **XX EURALEX International Congress** was held on **12–16 July 2022** in **Mannheim, Germany**. Themed "**Dictionaries and Society**", the conference brought together professional lexicographers, linguists, publishers, researchers, software developers and anyone interested in dictionaries and their educational, cultural, political and social impact in everyday life. Submissions on a wide range of topics were submitted, including:

- The Dictionary-Making Process
- Research on Dictionary Use
- Lexicography and Language Technologies
- Lexicography and Corpus Linguistics
- Bi- and Multilingual Lexicography
- Lexicography for Specialised Languages, Terminology and Terminography
- Lexicography of Lesser-Used and Under-Researched Languages
- Phraseology and Collocation
- Lexicography and Etymology
- Lexicological Issues of Lexicographical Relevance
- Reports on Lexicographical and Lexicological Projects

All submissions were reviewed in a double-blind peer review process by at least two members of the **Scientific Committee** (see page 14) for whose support we are very grateful. All decisions to accept or reject submissions for presentation at the congress and full papers for publication in the conference proceedings were based on the average score from reviews and in many cases on further evaluation by members of the **Programme Committee** (see page 14). We are very grateful to the EURALEX Board members who supported us as members of the Programme Committee, Iztok Kosem (Jožef Stefan Institute/University of Ljubljana, Slovenia), Robert Lew (Adam Mickiewicz University, Poland), Gilles-Maurice de Schryver (Ghent University, Belgium & University of Pretoria, South Africa), and Kristina Štrkalj Despot (Institute of Croatian Language and Linguistics, Zagreb, Croatia). Without the expertise and commitment of all colleagues who served on the Scientific and the Programme Committee, we would not have been able to maintain the high academic standard of presentations at EURALEX congresses and of their proceedings. Thank you!

This Book contains the full papers of keynotes, talks, posters, and software demonstrations of the XX. EURALEX International Congress, starting with the four **keynote papers** (**Part I**). We invited plenary speakers to address different aspects of our congress theme "Dictionaries and Society", such as the influence of society on lexicography, the role of women in lexicography, dictionary landscapes in multilingual societies, the role of dictionaries for language learners and traces of time and culture in (German) dictionaries. In this volume, Rufus Gouws (Stellenbosch University, South Africa), our 2022 Hornby Lecturer, discusses dictionaries as "bridges, dykes and sluice gates" in the multilingual society of South Africa. Thomas Gloning (University of Gießen, Germany) reflects on "Ways of living, communication and the dynamics of word usage". Nicola McLelland (University of Nottingham, UK) sheds new light on the role of women in German lexicography. Martina Nied Curcio (Università Roma Tre, Italy) explains which challenges for the use of dictionaries in language learning and teaching need to be overcome in the digital area.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Part II contains all other **full papers of talks, poster presentations, and software demonstrations** in thematic order (following an alphabetical order by their authors' surnames for each topic):

- Dictionaries and Society
- Lexicography: Status, Theory and Methods
- Corpora in Lexicography
- Data Models and Databases in Lexicography
- Dictionary Writing Systems and Lexicographic Tools
- Design and Publication of Dictionaries
- (Promoting) Dictionary Use
- Dictionary Projects
- Bilingual Dictionaries
- Specialised Dictionaries
- Historical Lexicography: German
- Historical Lexicography: Romance and Other Languages
- (Historical) Lexicology
- Neologisms and Lexicography
- Phraseology & Collocations
- Semantics

A total of sixty-seven full papers were accepted for publication. Of these, four papers were presented as part of the fourth edition of the Globalex Workshop on Lexicography and Neology (GWLN4; organised by Ilan Kernerman and Annette Klosa-Kückelhaus and integrated into EURALEX 2022 as an in-conference workshop on 15th July 2022).

An alphabetical index at the end of this publication contains all authors' names and facilitates finding papers by specific authors.

The Congress was organised by the Department of Lexical Studies ("Lexik") at the Leibniz Institute for the German Language (IDS) in Mannheim. Our sincere thanks go to all colleagues at IDS who supported the organisation of the congress and the publication of the abstract volume and, last but not least, the present conference proceedings. We would also like to thank all the sponsors (see page 13) who financially supported EURALEX 2022 and without whose generous support the congress could not have taken place.

As the chair of the XX EURALEX Organising Committee, I would like to gratefully acknowledge the support of the other members of our Organising Committee, Stefan Engelberg, Christine Möhrs, and Petra Storjohann, for their cooperation in the publication of this volume.

Annette Klosa-Kückelhaus
Chair of EURALEX 2022
June 2022

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

TABLE OF CONTENTS

Acknowledgements

Main Sponsors 13
Sponsors 13
Programme Committee 14
Scientific Committee 14

Part I: Overview on Keynotes, Talks, Posters, and Software Demonstrations

Keynotes 18
Talks 18
Posters 21
Software Demonstrations 21

Part II: Keynotes

Thomas Gloning
Ways of living, communication and the dynamics of word usage
How did German dictionaries cope with socio-cultural aspects and evolution of word usage and how could future systems do even better? 23

Rufus H. Gouws
Dictionaries: bridges, dykes, sluice gates 36

Nicola McLelland
Women in the history of lexicography
An overview, and the case of German 53

Martina Nied Curcio
Dictionaries, foreign language learners and teachers
New challenges in the digital era 71

Part III: Proceedings of Talks, Posters, and Software Demonstrations

Dictionaries and Society

Stefan Engelberg
Lexicography's entanglement with colonialism: The history of Tok Pisin lexicography as colonial history 87

Laura Giacomini/Paolo DiMuccio-Failla/Patrizio De Martin Pinter
The representation of culture-specific lexical items in monolingual learner's lexicography
The case of the electronic Phrase-Based Active Dictionaries 99

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Annette Klosa-Kückelhaus
Lexicography for society and with society – COVID-19 and dictionaries 113

Carolin Müller-Spitzer/Jan Oliver Rüdiger
The influence of the corpus on the representation of gender stereotypes in the dictionary. A case study of corpus-based dictionaries of German 129

Laura Pinnavaia
Identifying ideological strategies in the making of monolingual English language learner's dictionaries 142

Petra Storjohann
The public as linguistic authority: Why users turn to internet forums to differentiate between words 155

Lexicography: Status, Theory and Methods

Konan Kouassi
Mensch-Maschine-Interaktion im lexikographischen Prozess zu lexikalischen Informationssystemen 172

Ana Salgado/Rute Costa/Toma Tasovac
Applying terminological methods to lexicographic work: terms and their domains 181

Gilles-Maurice de Schryver
Metalexicography: an existential crisis 196

Corpora in Lexicography

Nils Diewald/Marc Kupietz/Harald Lüngen
Tokenizing on scale
Preprocessing large text corpora on the lexical and sentence level 208

Ana-Maria Gînsac/Mihai-Alex Moruz/Mădălina Ungureanu
The first Romanian dictionaries (17th century). Digital aligned corpus 222

Iztok Kosem
Trendi – a monitor corpus of Slovene 230

Simon Krek/Polona Gantar/Iztok Kosem
Extraction of collocations from the Gigafida 2.1 corpus of Slovene 240

Meike Meliss/Vanessa González Ribao
Vergleichbare Korpora für multilinguale kontrastive Studien
Herausforderungen und Desiderata 253

Irene Renau/Rogelio Nazar
Towards a multilingual dictionary of discourse markers
Automatic extraction of units from parallel corpus 262

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Chris A. Smith
Are phonesthemes evidence of a sublexical organising layer in the structure of the lexicon?
Testing the OED analysis of two phonesthemes with a corpus study of collocational behaviour of *sw-* and *fl-* words in the OEC 273

Data Models and Databases in Lexicography

Thierry Declerck
Integration of sign language lexical data in the OntoLex-Lemon framework 296

Birgit Füreder
Überlegungen zur Modellierung eines multilingualen ‚Periphrastikons':
Ein französisch-italienisch-spanisch-englisch-deutscher Versuch 301

David Lindemann/Penny Labropoulou/Christiane Klaes
Introducing LexMeta: a metadata model for lexical resources 310

Christian-Emil Smith Ore/Oddrun Grønvik/Trond Minde
Word banks, dictionaries and research results by the roadside 321

Ana Ostroški Anić/Ivana Brač
AirFrame
Mapping the field of aviation through semantic frames 334

Kristel Proost/Arne Zeschel/Frank Michaelis/Jan Oliver Rüdiger
MAP (Musterbank Argumentmarkierender Präpositionen)
A patternbank of argument-marking prepositions in German 346

Anna Vacalopoulou/Eleni Efthimiou/Stavroula-Evita Fotinea/Theodoros Goulas/Athanasia-Lida Dimou/Kiki Vasilaki
Organizing a bilingual lexicographic database with the use of WordNet 357

Dictionary Writing Systems and Lexicographic Tools

Nico Dorn
An automated cluster constructor for a narrated dictionary
The Cross-reference Clusters of *Wortgeschichte digital* 368

Mireille Ducassé/Archil Elizbarashvili
Finding lemmas in agglutinative and inflectional language dictionaries with logical information systems
The case of Georgian verbs 381

Velibor Ilić/Lenka Bajčetić/Snežana Petrović/Ana Španović
SCyDia – OCR for Serbian Cyrillic with diacritics 387

Dorielle Lonke/Ilan Kernerman/Vova Dzhuranyuk
Lexical data API 401

Takahiro Makino/Rei Miyata/Seo Sungwon/Satoshi Sato
Designing and building a Japanese controlled language for the automotive domain
Toward the development of a writing assistant tool 409

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Alberto Simões/Ana Salgado
Smart dictionary editing with LeXmart 423

Design and Publication of Dictionaries

Zita Hollós
Cross-Media-Publishing in der korpusgestützten Lernerlexikographie
Entstehung eines Lernerwörterbuchportals DaF 436

(Promoting) Dictionary Use

Andrea Abel
Wörterbücher der Zukunft in Bildungskontexten der Gegenwart
Eine Fallstudie aus dem Südtiroler Schulwesen 449

Carolina Flinz/Sabrina Ballestracci
Das LBC-Wörterbuch: Eine erste Benutzerstudie 460

Zoe Gavriilidou/Evi Konstandinidou
The effect of an explicit and integrated dictionary awareness intervention program on dictionary use strategies 471

Theresa Kruse/Ulrich Heid
Learning from students
On the design and usability of an e-dictionary of mathematical graph theory 480

Silga Sviķe
Survey analysis of dictionary-using skills and habits among translation students 494

Carole Tiberius/Jelena Kallas/Svetla Koeva/Margit Langemets/Iztok Kosem
An insight into lexicographic practices in Europe
Results of the extended ELEXIS Survey on User Needs 509

Agnes Wigestrand Hoftun
Consultation behavior in L1 error correction
An exploratory study on the use of online resources in the Norwegian context 522

Dictionary Projects

Hauke Bartels
The long road to a historical dictionary of Lower Sorbian
Towards a lexical information system 540

Polona Gantar/Simon Krek
Creating the lexicon of multi-word expressions for Slovene
Methodology and structure 549

Zoe Gavriilidou/Apostolos Garoufos
The lexicographic protocol of Mikaela_Lex: A free online school dictionary of Greek accessible for visually-impaired senior elementary children 563

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Vanessa González Ribao
Fachlexikografie in digitalem Zeitalter
Ein metalexikografisches Forschungsprojekt 569

Peter Meyer
Lehnwortportal Deutsch: a new architecture for resources on lexical borrowings 578

Iryna Ostapova/Volodymyr Shyrokov/Yevhen Kupriianov/Mykyta Yablochkov
Etymological dictionary in digital environment 584

Anna Pavlova
Mehrsprachige Datenbank der Phrasem-Konstruktionen 594

Ralf Plate
Word Families in Diachrony
An epoch-spanning structure for the word families of older German 605

Kyriaki Salveridou/Zoe Gavriilidou
Compilation of an Ancient Greek – Modern Greek online thesaurus for teaching purposes: microstructure and macrostructure 614

Bilingual Dictionaries

Voula Giouli/Anna Vacalopoulou/Nikos Sidiropoulos/Christina Flouda/Athanasios Doupas/Gregory Stainhaouer
From mythos to logos: A bilingual thesaurus tailored to meet users' needs within the ecosystem of cultural tourism 625

Anke Müller/Gabriele Langer/Felicitas Otte/Sabrina Wähl
Creating a dictionary of a signed minority language
A bilingualized monolingual dictionary of German Sign Language 635

Specialized Dictionaries

Maria Aldea
Bien écrire, bien parler au XIXe siècle. Le rôle du dictionnaire dans l'apprentissage de la langue maternelle: Le cas du roumain 650

Harald Bichlmeier
Almanca tuhfe/Deutsches Geschenk (1916)
oder: Wie schreibt man deutsch mit arabischen Buchstaben? 660

María Pozzi
Design of a dictionary to help school children to understand basic mathematical concepts 678

Stefan J. Schierholz/Monika Bielinska/Maria José Domínguez Vázquez/Rufus H. Gouws/Martina Nied Curcio
The EMLex Dictionary of Lexicography (EMLexDictoL) 690

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Historical Lexicography: German

Volker Harm
Wortgeschichte digital: A historical dictionary of New High German 701

Andrea Moshövel
Skatologischer Wortschatz im Frühneuhochdeutschen als kulturgeschichtliche und lexikographische Herausforderung 711

Historical Lexicography: Romance and Other Languages

Maria Arapopoulou/Georgios Kalafikis/Dimitra Karamitsou/Efstratios Sarischoulis/Sotiris Tselikas
"Vocabula Grammatica": threading a digital Ariadne's String in the labyrinth of Ancient Greek scholarship 725

Anaïs Chambat
La lignée «Capuron-Nysten-Littré» entre ruptures et continuités doctrinales 735

Mihai-Alex Moruz/Mădălina Ungureanu
17th-century Romanian lexical resources and their Influence on Romanian written tradition 745

Clarissa Stincone
Usage labels in Basnage's *Dictionnaire universel* (1701) 755

Marija Žarković
The legal lexicon in the first dictionary of the Spanish Royal Academy (1726–1739)
The Concept of the Judge 765

(Historical) Lexicology

Ellert Thor Johannsson
Old words and obsolete meanings in Modern Icelandic 777

Pius ten Hacken/Renáta Panocová
The etymology of internationalisms
Evidence from German and Slovak 792

Neologisms and Lexicography

Ieda Maria Alves/Bruno Maroneze
From society to neology and lexicography
Relationships between morphology and dictionaries 804

Jun Choi/Hae-Yun Jung
On loans in Korean new word formation and in lexicography 814

Lars Trap-Jensen/Henrik Lorentzen
Recent neologisms provoked by COVID-19 in the Danish Language and in The Danish Dictionary 825

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Gilles-Maurice de Schryver/Minah Nabirye
Towards a monitor corpus for a Bantu language
A case study of neology detection in Lusoga 833

Phraseology & Collocations

Maria Ermakova/Alexander Geyken/Lothar Lemnitzer/Bernhard Roll
Integration of multi-word expressions into the Digital Dictionary of German Language (DWDS)
Towards a lexicographic representation of phraseological variation 851

Semantics

Robert Krovetz
An investigation of sense ordering across dictionaries with respect to lexical semantic relationships 862

Index of Authors

Index of Authors 871

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Acknowledgements

We would like to thank all those who have supported the XX EURALEX International Conference financially:

Main Sponsors

Funded by DFG Deutsche Forschungsgemeinschaft German Research Foundation

Hornby A. S. Hornby Educational Trust

elexis european lexicographic infrastructure

IDS LEIBNIZ-INSTITUT FÜR DEUTSCHE SPRACHE

Freunde des Leibniz-Instituts für Deutsche Sprache e.V.

Sponsors

BLOOMSBURY ACADEMIC

DE GRUYTER

DUDEN

ESV ERICH SCHMIDT VERLAG

OED

BUSKE

Frank & Timme

lexicala BY K DICTIONARIES

Universitätsverlag WINTER Heidelberg

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

We would like to thank all those who have contributed to reviewing the submissions and papers:

Programme Committee

Gilles-Maurice **de Schryver** (Ghent University, Belgium & University of Pretoria, South Africa)

Stefan **Engelberg** (The Leibniz Institute for the German Language, Germany)

Annette **Klosa-Kückelhaus** (The Leibniz Institute for the German Language, Germany)

Iztok **Kosem** (Jožef Stefan Institute / University of Ljubljana, Slovenia)

Robert **Lew** (Adam Mickiewicz University, Poland)

Christine **Möhrs** (The Leibniz Institute for the German Language, Germany)

Petra **Storjohann** (The Leibniz Institute for the German Language, Germany)

Kristina **Štrkalj Despot** (Institute of Croatian Language and Linguistics, Croatia)

Scientific Committee

Andrea **Abel** (EURAC, Italy)

Arleta **Adamska-Sałaciak** (Adam Mickiewicz University, Poland)

Hauke **Bartels** (Sorbian Institute, Germany)

Hans **Bickel** (Schweizerisches Idiotikon, Switzerland)

Anna **Braasch** (University of Copenhagen, Denmark)

Dominik **Brückner** (The Leibniz Institute for the German Language, Germany)

Thomas **Burch** (Trier Center for Digital Humanities, Germany)

Lut **Colman** (Dutch Language Institute, Netherlands)

Paul **Cook** (University of New Brunswick, Canada)

Gilles-Maurice **de Schryver** (Ghent University, Belgium & University of Pretoria, South Africa)

Janet **DeCesaris** (Pompeu Fabra University, Spain)

Idalete Maria Silva **Dias** (University of Minho, Portugal)

María José **Domínguez Vázquez** (University of Santiago de Compostela, Spain)

Philip **Durkin** (Oxford University Press, Great Britain)

Anne **Dykstra** (Fryske Academy, Netherlands)

Anna **Dziemianko** (Adam Mickiewicz University, Poland)

Ilse **Feinauer** (Stellenbosch University, South Africa)

Edward **Finegan** (University of Southern California, USA)

Carolina **Flinz** (University of Milan, Italy)

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Thierry **Fontenelle** (European Investment Bank, Belgium)

Polona **Gantar** (University of Ljubljana, Slovenia)

Zoe **Gavriilidou** (Democritus University of Thrace, Greece)

Alexander **Geyken** (Berlin-Brandenburg Academy of Sciences, Germany)

Sylviane **Granger** (Catholic University of Louvain, Belgium)

Oddrun **Grønvik** (University of Oslo, Norway)

Volker **Harm** (The Göttingen Academy of Sciences and Humanities, Germany)

Ulrich **Heid** (Hildesheim University, Germany)

Zita **Hollós** (Károli Gáspár University, Hungary)

Miloš **Jakubíček** (Lexical Computing CZ s.r.o., Czech Republic)

Maarten **Janssen** (University of Vienna, Austria)

Besim **Kabashi** (Friedrich-Alexander University Erlangen, Germany)

Jelena **Kallas** (Institute of the Estonian Language, Estonia)

Heidrun **Kämper** (The Leibniz Institute for the German Language, Germany)

Ilan **Kernerman** (K Dictionaries, Israel)

Alexander **Koplenig** (The Leibniz Institute for the German Language, Germany)

Iztok **Kosem** (Jožef Stefan Institute / University of Ljubljana, Slovenia)

Simon **Krek** (Jožef Stefan Institute / University of Ljubljana, Slovenia)

Tanara Zingano **Kuhn** (University of Coimbra, Portugal)

Kathrin **Kunkel-Razum** (Duden-Verlag, Germany)

Margit **Langemets** (Institute of the Estonian Language, Estonia)

Lothar **Lemnitzer** (Berlin-Brandenburg Academy of Sciences, Germany)

Robert **Lew** (Adam Mickiewicz University, Poland)

Marie-Claude **L'Homme** (University of Montreal, Canada)

Anja **Lobenstein-Reichmann** (The Göttingen Academy of Sciences and Humanities, Germany)

Henrik **Lorentzen** (The Danish Language and Literature Society, Denmark)

Carla **Marello** (University of Turin, Italy)

Tinatin **Margalitadze** (Ilia State University, Georgia)

John P. **McCrae** (National University of Ireland, Ireland)

Peter **Meyer** (The Leibniz Institute for the German Language, Germany)

Frank **Michaelis** (The Leibniz Institute for the German Language, Germany)

Julia **Miller** (University of Adelaide, Australia)

Fabio **Mollica** (University of Milan, Italy)

Orion **Montoya** (Brandeis University, USA)

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Rosamund **Moon** (University of Birmingham, Great Britain)

Carolin **Müller-Spitzer** (The Leibniz Institute for the German Language, Germany)

Kilim **Nam** (Kyungpook National University, South Korea)

Hilary **Nesi** (Coventry University, Great Britain)

Vincent **Ooi** (National University of Singapore, Singapore)

Maike **Park** (The Leibniz Institute for the German Language, Germany)

Ralf **Plate** (The Academy of Sciences and Literature Mainz / University of Trier, Germany)

Kristel **Proost** (The Leibniz Institute for the German Language, Germany)

Natascia **Ralli** (EURAC, Italy)

Stefan J. **Schierholz** (Friedrich-Alexander University Erlangen, Germany)

Thomas **Schmidt** (The Leibniz Institute for the German Language, Germany)

Hindrik **Sijens** (Fryske Academy, Netherlands)

Egon W. **Stemle** (EURAC, Italy)

Frieda **Steurs** (Dutch Language Institute, Netherlands)

Kathrin **Steyer** (The Leibniz Institute for the German Language, Germany)

Philipp **Stöckle** (Austrian Academy of Sciences, Austria)

Kristina **Štrkalj Despot** (Institute of Croatian Language and Linguistics, Croatia)

Janusz **Taborek** (Adam Mickiewicz University, Poland)

Elsabé **Taljard** (University of Pretoria, South Africa)

Pius **ten Hacken** (University of Innsbruck, Austria)

Carole **Tiberius** (Dutch Language Institute, Netherlands)

Yukio **Tono** (Tokyo University of Foreign Studies, Japan)

Lars **Trap-Jensen** (The Danish Language and Literature Society, Denmark)

Anna **Vacalopoulou** (Institute for Language and Speech Processing, Greece)

Carlos **Valcárcel Riveiro** (University of Vigo, Spain)

Ruth **Vatvedt Fjeld** (University of Oslo, Norway)

Craig **Volker** (James Cook University Cairns, Australia)

Sabine **Wahl** (Austrian Academy of Sciences, Austria)

Geoffrey **Williams** (Université Bretagne Sud, France)

Sascha **Wolfer** (The Leibniz Institute for the German Language, Germany)

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Part I: Overview on Keynotes, Talks, Posters, and Software Demonstrations

Keynotes

Thomas Gloning: Ways of living, communication and the dynamics of word usage. How did German dictionaries cope with socio-cultural aspects and evolution of word usage and how could future systems do even better?

Rufus H. Gouws: Dictionaries: bridges, dykes and sluice gates

Nicola McLelland: Women in the history of lexicography. An overview, and the case of German

Martina Nied Curcio: Dictionaries, foreign language learners and teachers. New challenges in the digital era

Talks

Andrea Abel: Wörterbücher der Zukunft in Bildungskontexten der Gegenwart. Eine Fallstudie aus dem Südtiroler Schulwesen

Maria Aldea: Bien écrire, bien parler au XIXe siècle. Le rôle du dictionnaire dans l'apprentissage de la langue maternelle. Le cas du roumain

Ieda Maria Alves/Bruno Maroneze: From society to neology and lexicography. Relationships between morphology and dictionaries

Maria Arapopoulou/Georgios Kalafikis/Dimitra Karamitsou/Efstratios Sarischoulis/Sotiris Tselika: "Vocabula Grammatica": threading a digital Ariadne's String in the labyrinth of Ancient Greek scholarship

Hauke Bartels: The long road to a historical dictionary of Lower Sorbian. Towards a lexical information system

Harald Bichlmeier: *Almanca tuhfe / Deutsches Geschenk* (1916) oder: Wie schreibt man deutsch mit arabischen Buchstaben?

Anaïs Chambat: La lignée «Capuron-Nysten-Littré» entre ruptures et continuités doctrinales

Jun Choi/Hae-Yun Jung: On loans in Korean new word formation and in lexicography

Gilles-Maurice de Schryver: Metalexicography: an existential crisis

Gilles-Maurice de Schryver/Minah Nabirye: Towards a monitor corpus for a Bantu language. A case study of neology detection in Lusoga

Stefan Engelberg: Lexicography's entanglement with colonialism: The history of Tok Pisin lexicography as colonial history

Maria Ermakova/Alexander Geyken/Lothar Lemnitzer/Bernhard Roll: Integration of multi-word expressions into the Digital Dictionary of German Language (DWDS). Towards a lexicographic representation of phraseological variation

Carolina Flinz/Sabrina Ballestracci: Das LBC-Wörterbuch: Eine erste Benutzerstudie

Polona Gantar/Simon Krek: Creating the lexicon of multi-word expressions for Slovene Methodology and structure

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Zoe Gavriilidou/Evi Konstandinidou: The effect of an explicit and integrated dictionary awareness intervention program on dictionary use strategies

Laura Giacomini/Paolo DiMuccio-Failla/Patrizio De Martin Pinter: The representation of culture-specific lexical items in monolingual learner's lexicography

Voula Giouli/Anna Vacalopoulou/Nikos Sidiropoulos/Christina Flouda/Athanasios Doupa/Gregory Stainhaouer: From mythos to logos: A bilingual thesaurus tailored to meet users' needs within the ecosystem of cultural tourism

Volker Harm: *Wortgeschichte digital*: A historical dictionary of New High German

Zita Hollós: Cross-Media-Publishing in der korpusgestützten Lernerlexikographie. Entstehung eines Lernerwörterbuchportals DaF

Ellert Thor Johannsson: Old words and obsolete meanings in Modern Icelandic

Annette Klosa-Kückelhaus: Lexicography for society and with society – COVID-19 and dictionaries

Iztok Kosem: Trendi – a monitor corpus of Slovene

Konan Kouassi: Mensch-Maschine-Interaktion im lexikographischen Prozess zu lexikalischen Informationssystemen

Simon Krek/Polona Ganta/Iztok Kosem: Extraction of collocations from the Gigafida 2.1 corpus of Slovene

Robert Krovetz: An investigation of sense ordering across dictionaries with respect to lexical semantic relationships

Theresa Kruse/Ulrich Heid: Learning from students. On the design and usability of an e-dictionary of mathematical graph theory

David Lindemann/Penny Labropoulou/Christiane Klaes: Introducing LexMeta: a metadata model for lexical resources

Takahiro Makino/Rei Miyata/Seo Sungwon/Satoshi Sato: Designing and building a Japanese controlled language for automotive domain. Toward the development of a writing assistant tool

Mihai-Alex Moruz/Mădălina Ungureanu: 17th-century Romanian lexical resources and their Influence on Romanian written tradition

Andrea Moshövel: Skatologischer Wortschatz im Frühneuhochdeutschen als kulturgeschichtliche und lexikographische Herausforderung

Anke Müller/Gabriele Langer/Felicitas Otte/Sabrina Wähl: Creating a dictionary of a signed minority language: A bilingualized monolingual dictionary of German Sign Language

Carolin Müller-Spitzer/Jan Oliver Rüdiger: The influence of the corpus on the representation of gender stereotypes in the dictionary. A case study of corpus-based dictionaries of German

Iryna Ostapova/Volodymyr Shyrokov/Yevhen Kupriianov/Mykyta Yablochlov: Etymological dictionary in digital environment

Ana Ostroški Anić/Ivana Brač: AirFrame. Mapping the field of aviation through semantic frames

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Anna Pavlova: Mehrsprachige Datenbank der Phrasem-Konstruktionen

Laura Pinnavaia: Identifying ideological strategies in the making of monolingual English language learner's dictionaries

Ralf Plate: Word families in diachrony. An epoch-spanning structure for the word families of older German

María Pozzi: Design of a dictionary to help school children to understand basic mathematical concepts

Kristel Proost/Arne Zeschel/Frank Michaelis/Jan Oliver Rüdiger: MAP (Musterbank Argumentmarkierender Präpositionen). A patternbank of argument-marking prepositions in German

Irene Renau/Rogelio Nazar: Towards a multilingual dictionary of discourse markers. Automatic extraction of units from parallel corpus

Ana Salgado/Rute Costa/Toma Tasovac: Applying terminological methods to lexicographic work: terms and their domains

Kyriaki Salveridou/Zoe Gavriilidou: Compilation of an Ancient Greek – Modern Greek online thesaurus for teaching purposes: microstructure and macrostructure

Stefan J. Schierholz/Monika Bielinska/Maria José Domínguez Vázquez/Rufus H. Gouws/Martina Nied Curcio: The EMLex Dictionary of Lexicography (EMLexDictoL)

Alberto Simões/Ana Salgado: Smart dictionary editing with LeXmart

Christian-Emil Smith Ore/Oddrun Grønvik/Trond Minde: Word banks, dictionaries and research results by the roadside

Clarissa Stincone: Usage labels in Basnage's *Dictionnaire universel* (1701)

Petra Storjohann: The public as linguistic authority: Why users turn to internet forums to differentiate between words

Pius ten Hacken/Renáta Panocová: The etymology of internationalisms. Evidence from German and Slovak

Carole Tiberius/Jelena Kallas/Svetla Koeva/Margit Langemets/Iztok Kosem: An insight into lexicographic practices in Europe. Results of the extended ELEXIS Survey on User Needs

Lars Trap-Jensen/Henrik Lorentzen: Recent neologisms provoked by COVID-19 in the Danish language and in *The Danish Dictionary*

Anna Vacalopoulou/Eleni Efthimiou/Stavroula-Evita Fotinea/Theodoros Goulas/Athanasia-Lida Dimou/Kiki Vasilaki: Organizing a bilingual lexicographic database with the use of WordNet

Agnes Wigestrand Hoftun: Consultation behavior in L1 error correction. An exploratory study on the use of online resources in the Norwegian context

Marija Žarković: The legal lexicon in the first dictionary of the Spanish Royal Academy (1726–1739). The Concept of the Judge

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Posters

Thierry Declerck: Integration of sign language lexical data in the OntoLex-Lemon framework

Nils Diewald/Marc Kupietz/Harald Lüngen: Tokenizing on scale. Preprocessing large text corpora on the lexical and sentence level

Birgit Füreder: Überlegungen zur Modellierung eines multilingualen ‚Periphrastikons'. Ein französisch-italienisch-spanisch-englisch-deutscher Versuch

Zoe Gavriilidou/Apostolos Garoufos: The lexicographic protocol of MikaelaLex. A free online school dictionary of Greek accessible for visually-impaired senior elementary children

Ana-Maria Gînsac/Mihai-Alex Moruz/Mădălina Ungureanu: The first Romanian dictionaries (17th century). Digital aligned corpus

Vanessa Gonzalez Ribao: Fachlexikografie in digitalem Zeitalter: Ein metalexikografisches Forschungsprojekt

Velibor Ilić/Lenka Bajčetić/Snežana Petrović/Ana Španović: SCyDia – OCR for Serbian Cyrillic with Diacritics

Meike Meliss/Vanessa González Ribao: Vergleichbare Korpora für multilinguale kontrastive Studien. Herausforderungen und Desiderata

Chris A. Smith: Are phonesthemes evidence of a sublexical organising layer in the structure of the lexicon? Testing the OED analysis of two phonesthemes with a corpus study of collocational behaviour of *sw-* and *fl-* words in the OEC

Silga Sviķe: Survey analysis of dictionary-using skills and habits among translation students

Software Demonstrations

Nico Dorn: An automated cluster constructor for a narrated dictionary. The Cross-reference Clusters of *Wortgeschichte digital*

Mireille Ducassé/Archil Elizbarashvili: Finding lemmas in agglutinative and inflectional language dictionaries with logical information systems: The case of Georgian verbs

Dorielle Lonke/Ilan Kernerman/Vova Dzhuranyuk: Lexical data API

Peter Meyer: Lehnwortportal Deutsch: a new architecture for resources on lexical borrowings

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Part II: Keynotes

Thomas Gloning

WAYS OF LIVING, COMMUNICATION AND THE DYNAMICS OF WORD USAGE

How did German dictionaries cope with socio-cultural aspects and evolution of word usage and how could future systems do even better?

Abstract Words and their usages are in many cases closely related to or embedded in social, cultural, technical and ideological contexts. This does not only apply to individual words and specific senses, but to many vocabulary zones as well. Moreover, the development of words is often related to aspects of socio-cultural evolution in a broad sense. In this paper I will have a look at traditional dictionaries and digital lexical systems focussing on the question how they deal with socio-cultural and discourse-related aspects of word usage. I will also propose a number of suggestions how future digital lexical systems might be enriched in this respect.

Keywords Digital lexical systems; extended search; vocabulary organization in dictionaries; forms of representation in digital lexicography

1. Introduction

Word usages and their development as well as the organization of the lexicon and its dynamics are both intimately related to aspects of social and cultural evolution in a broad sense. Institutional structures, the evolvement of topics, cultural contacts, new technologies etc. are mirrored in a complex way in our modes of communication and the words we use thereby. Considering the relationship between word usage and this complex architecture of multiple aspects of ways of living and given the historical evolution of almost all aspects of this interrelation, one can ask:

- How did traditional dictionaries of German cope with both the structure and the evolution of word usages in their socio-cultural settings?
- In this respect, what might be fruitful perspectives for future digital lexical systems?

In this paper, I should like to proceed in three steps: First, I should like to give an overview of the interrelatedness of forms of living ("culture"), word usage and their evolution over time. I will discuss this aspect using examples from both modern times and the history of the German language (section 2). Secondly, I should like to demonstrate how different aspects of culture and cultural evolution have been treated (or have *not* been treated) in dictionaries of German (section 3). Thirdly, I should like to give examples and propose suggestions of how future digital systems might improve the documentation and the analysis of word usage and lexical groups in the context of culture and cultural evolution (section 4).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

2. Forms of living ("culture") and word usage: interrelations and evolution over time

Word usage and its dynamics are intimately related to social, political, cultural and intellectual structures and to changes in all major aspects of ways of living (Wittgenstein's "Lebensform", 'form of life').[1] This is true not only for individual words but also for "zones" of vocabulary that can be grouped according to different socio-cultural and communicative criteria. Take, e. g., nutrition, health, sports, sex, technology, the "ecology" of ideas, controversial public topics like immigration, abortion, nuclear energy or climate change, views on beauty or health, tourism, economic developments, the world of labour, aspects of gender and the relation of men, women and persons with other identities, forms of clothing, the things we use in everyday life and so on. Moreover, there are words used for specific communicative functions, e. g. evaluation in film reviews (*grandios* 'grandiose, magnificent') or means of politeness (*freundlicherweise* 'kindly, please', *gefälligst* 'please, by courtesy'), which may change over time. And then there is a plethora of traditional lexicological aspects like word formation, the spatial profile of words, the question of the history of foreign words, semantic relations (e. g. *Vetter* and *Cousin* 'cousin'), words that are adopted from languages for specific purposes, etc. Most of these lexicological aspects and their dynamics can have a "cultural" background as well.

I will now illustrate these points with several examples for relations between the socio-cultural and the lexical. Some are fairly obvious, others might not be trivial, some are meant to show that the aspects of the socio-cultural and the communicative are intertwined in an intricate way.

The first example highlights the vocabulary of public administration. In the annual report 2015 of the psychiatric clinic for children and adolescents of the "Kanton" Zurich, Switzerland, there is a preface, which was signed:[2]

(1) **Regierungsrat Dr. Thomas Heiniger**
Gesundheitsdirektor Kanton Zürich

A common understanding of the key expressions in this signature is that *Regierungsrat* is an expression for a certain rank in a system of public administration and that *Gesundheitsdirektor* is used to refer to a specific function in the medical sector of Swiss public administration. Moreover, we can infer that the role of a Swiss Gesundheitsdirektor is somehow related to the political unit "Kanton" in Switzerland. What readers not familiar with the Swiss administration and its political system probably do not know: What exactly is the position of this rank in the system of ranks? What exactly are the function, the rights and duties, etc. of a *Gesundheitsdirektor* and how does the structure of the surrounding medical administration look like in which this role is embedded? Obviously, contributing prefaces is one of his/her duties, but what else is he or she responsible for? The question "What is an X?" is closely related to "How is the word *X* used?" In a lexical system, we would probably not expect to find all answers to the abovementioned questions, but we expect at least basic information

[1] "Das Wort ‚*Sprachspiel*' soll hier hervorheben, daß das Sprechen der Sprache ein Teil ist einer Tätigkeit, oder einer Lebensform" (Wittgenstein 1969, pp. 300 = PU §23); (The word *Sprachspiel*, 'language game' here is meant to emphasize that the speaking of language is a part of an activity, or a *Lebensform*, a way of life). For a systematic exposition cf. Hacker (2011).

[2] *Regierungsrat* refers to a high official in public administration; *Gesundheitsdirektor* refers to the chief official of the public health administration in a Swiss Canton.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

on what a *Gesundheitsdirektor* is and how the word was used in space, time, different text types, etc. Consider a further example, Uwe Johnson's "Mutmassungen über Jakob" (1959): In speaking about the (dead) protagonist's profession, the author uses the word *Dispatcher*. Today's reader might want to know whether this was Eastern German or if this word was just from some earlier era of railroad transportation. This kind of information can be crucial in analyzing the lexical profile of literary works and how they relate to social and cultural aspects. A further example from administrative language in a literary text is Musil's use of *Sektionschef*, which poses similar questions related to early 20th century Austria.

Broadening the view from single words to vocabulary zones, we can ask for the vocabulary related to public administration and where we might find it. One of the German language sources for this field is the periodical "Behörden-Spiegel", which is devoted to a broad range of topics relevant for the organization of public affairs. The journal is available both on its website and in pdf format on issuu.com. Usually, the work and the language use of public administration does not appear on the radar of citizens as long as things run smoothly. However, making things run smoothly in the public sphere requires a lot of communication. Looking at one issue of the "Behörden-Spiegel" (2022–04), we find words like the following, which for a start I present in an unordered list as they came along while reading:

(2) *fahrradfreundlich*, *netzwerkfähig*, *Digitalisierung*, *Resilienz*, *Sperrmüll*, *Multimobilität*, *Geschäftsbereichsleiter*, *Teamleiterin*, *zeitnah*, *Beschwerdeverfahren*, *Open Source*, *Best Practices*, *Bürgerdienste*, *Videokommunikation*, *Chief Executive Officer*, *CEO*, *Cyber-Angriff*, *repräsentative Umfrage*, *Steuerbetrügereien*, *Steueridentifikationsnummer*, *Dienstunfall*, *Arbeitsunfall*, *Dienststelle*, *Changemanagement*, *Fallbearbeitung*, *Nachrichtendienst*, *Verwaltungsvereinbarung*, *Cloud*, *Interimslösung*, *Großschadenlage*, *Einsatzkommunikation*, *Transportflugzeug*, *Transformation*, *desolat*.[3]

On closer inspection, we recognize certain internal lexical fields, e. g. for topics around the digital (*Open Source*), for aspects of the internal organization of public administration (*Changemanagement*), in respect of different task sectors like civil security and protection (*Großschadenslage*, *Cyber-Angriff*) or tax administration (*Steueridentifikationsnummer*) and also expressions for goals (*Multimobilität*). Most of these and other fields have a long history, e. g. the social organization of unemployment. Consequently, there is a textual cosmos and a specific vocabulary sector related to this topic since the second half of the 19th century. Obviously, dictionaries are not ideal instruments for analyzing the structure or narrating the history of whole vocabulary zones, but they may serve as the place for the documentation of the words that are part of a structural analysis or a historical narrative in some kind of monographic form.[4]

I now turn to a second example for the relatedness between word usage and socio-cultural aspects, which highlights the connection between word usage and public debates about

[3] DeepL provides the following equivalents: *bike-friendly*, *network-enabled*, *digitization*, *resilience*, *bulky waste*, *multi-mobility*, *business unit manager*, *team leader*, *timely*, *grievance process*, *open source*, *best practices*, *citizen services*, *video communications*, *chief executive officer*, *CEO*, *cyber attack*, *representative survey*, *tax fraud*, *tax identification number*, *service accident*, *work accident*, *office*, *change management*, *case management*, *intelligence*, *stewardship agreement*, *cloud*, *interim solution*, *major incident*, *mission communications*, *transport aircraft*, *transformation*, *desolate*.

[4] Cf. the combination of monograph and dictionary component in the discourse projects of Heidrun Kämper, e. g. 2012 (monograph) and 2013 (dictionary) on the German discourse on democracy in the late 1960s. An earlier example for such a combination is Schirmer's (1911) historical dictionary and lexicological monograph ("mit einer systematischen Einleitung") on the language and vocabulary of merchants.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

controversial topics. It is one of the crucial assumptions in the linguistic study of discourses on public topics that the use of specific words plays a crucial role in formulating and supporting positions, in propagating views and in following specific goals in a controversy. The team around Georg Stötzel and Martin Wengeler (e. g. 1995) has provided many studies on forms of discussing controversial topics in post-WWII history like military rearmament, immigration, nuclear energy, sexuality or abortion. Heidrun Kämper conducted four projects on word usage in times of change,[5] thereby developing a specific format with an organized combination of a monography and a discourse dictionary. The dictionary components are available within the IDS's lexical portal (www.owid.de).

While the methodology of investigation on word usage in controversial discourse is well established and we have quite a number of good case studies, there is still a lot of work to be done regarding the discourse characteristics of many words. E. g., at present, there are no fully developed dictionary entries for the word *Generationengerechtigkeit* ('intergenerational equity') which is a core word in a number of controversial topics.[6] The open DWDS corpora provide more than 500 instances, the corpora at the IDS and the DWDS, that are available after registration provide several thousands of instances. In addition, there are Google results, which are not quantifiable in a reliable way. More importantly, Urban/Ehlscheid (2020, p. 25) write in "Aus Politik und Zeitgeschichte", an important periodical on political research and education:

> „Generationengerechtigkeit" ist eine Kernvokabel der zeitgenössischen Debatte über Gegenwart und Zukunft des Sozialstaates. Begriffe wie „Generationensolidarität" und „Generationenkrieg" bilden einen Rahmen, innerhalb dessen sich eine Vielzahl von Deutungen der Generationenverhältnisse bewegt. Dabei stechen der Generationen- und der Gerechtigkeitsbegriff selbst in den an umstrittenen Begriffen reichen Sozialwissenschaften und der Philosophie hervor.[7]

In order to write a fully-fledged article on discourse vocabulary like *Generationengerechtigkeit* it takes more than giving a definition and it is an open question how the suggestions of discourse dictionaries or narrative lexicography can be adopted for general dictionaries or "all purpose" digital lexical systems (see section 4).

My third example comes from the arts; it highlights several aspects of the German vocabulary of jazz and its evolution (cf. Gloning 2022). Roughly speaking, the history of jazz music in the German speaking countries began in the 1920s. While some were enthusiastic about jazz, it was also the object of acidic verdicts (*Negermusik*). Since the 1950s new styles of playing evolved with new loan words like *Cool Jazz*, *Bebop* or *Fusion*. In many cities an infrastructure was developed (*Jazzkeller*, *Jazzclub*, *Jazz matinée*). But most importantly, the artistic and social developments were accompanied by a growing body of texts, e. g. a dedicated journal like the "Jazz Podium" or the "Jazz-Buch" (1957) by Joachim Ernst Berendt, one of the first of early overall presentations of jazz in German. The systematic organization of

5 Demokratiediskurs 1918–25; Schulddiskurs 1945–55; Protestdiskurs 1967/68; Schlüsselwörter 1989/90.

6 The article "Generationengerechtigkeit" in Wikipedia includes information about discourse developments and its chronology but not a systematic documentation of word usage in German. Cf. Deutsch (2022) for reflexions on the role of specialized lexicography in times of Google and Wikipedia.

7 DeepL translates: *Generational justice* is a core vocabulary of the contemporary debate on the present and future of the welfare state. Terms such as *generational solidarity* and *generational war* provide a framework within which a variety of interpretations of generational relations operate. In this context, the concepts of generation and justice stand out even in the social sciences and philosophy, which are rich in controversial terms.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Berendt's book comes with a rich body of vocabulary items, which mirrors the "system" of jazz and its positions, e. g. expressions for kinds of musicians (*Tenorsaxofonist*), for instruments (*Klarinette*), for genres (*Ballade*) or aspects of musical theory (*Offbeat*). Some of these words, e. g. *Klarinette*, are not restricted to the field of jazz, but they are indispensable when writing about jazz, other words like *Ballade* have jazz-specific senses. Again, the role of loan words, foreign words and of word formation becomes evident for the development of jazz vocabulary. Radio and TV broadcasts, feuilleton articles and specialized websites equally contribute to a rich communicative cosmos of jazz with a broad and well-organized vocabulary. Similarly, these kinds of interrelations between artistic, social, textual, medial and lexical developments can be studied in other fields as well (e. g. theatre, painting, dance). E. g., the praxis of artistic dancing has changed and this was mirrored in the history of its vocabulary. A word like *Ausdruckstanz* ('expressive dance, expression dance') refers to one of the 'new' directions of 20th century artistic dance. The word *Laban-Notation* or *Labanotation* was used for a specific notation system for dance movements, invented by Rudolf von Laban in the 1920s. The history of a cultural field is also the history of its vocabulary.

Other examples could be given from technology, medicine, politics, military and military technology, sports, gender topics, beauty ideals, sexuality, food and nutrition, tourism, knowledge systems of all kinds, and many others in respect of both their current state as well as their historical manifestations. They all would show that almost all aspects of our culture, our "way of living" are intricately related to word usage, but in different ways. In addition, many of these fields show an evolutionary dynamic regarding the relation of expert usages and the use in more popular contexts.

3. Word usage, culture and cultural evolution in dictionaries of German

I shall now have a look at some of the dictionaries and digital lexical systems of German and ask how they deal with these social, cultural and discourse related aspects of word usage. Here, one can take two different perspectives.

a) First, one can start from the dictionaries and ask what they (can) contribute in respect of the relation of word usage to social and cultural aspects, also in an historical perspective.
b) Secondly, one can look from specific words, senses or vocabulary zones and ask what different dictionaries (can) contribute to their description and to the documentation of their social and cultural aspects.

Looking from either perspective, there are four questions: 1) (How) Are words and senses pertaining to social, cultural and discourse developments documented in German dictionaries? 2) Are the social, cultural, discourse-related aspects of the use of words part of their description and documentation? 3) Do we get information on the interrelations between words within specific cultural, social or discourse-related vocabulary zones? 4) Are there ways of searching for or addressing specific vocabulary zones and the words that are their elements?

However, in answering such questions we must be aware of what we cannot reasonably expect from dictionaries or from digital lexical systems in their current form. This caveat will lead to the question what might be new features of next generation systems (section 4).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

3.1 (How) Are words and senses pertaining to social, cultural and discourse developments documented in German dictionaries?

It is obvious that in this article I cannot answer such a complex question in a straightforward and exhaustive manner. Therefore, I should like to mention only some aspects that are pertinent to this question.

One of the most important questions for users is: Do I find a word I am looking for, e. g. the word *Generationengerechtigkeit*? Now it is easy and cheap to say that certain words are missing in dictionaries and digital lexical systems. Nevertheless, I have to say that as of May 2022 I did not find the word in OWID, in DWDS nor in other sources. However, if one does not find a specific word in the dictionary components, the digital lexical systems at the IDS and the BBAW will at least provide corpus findings that allow users to investigate the word usage her- or himself. Furthermore, in a digital environment it is possible to add new articles on an (almost) daily basis. Interactive web elements can invite users to propose new articles they are interested in. If we compare the situation with the "Deutsches Wörterbuch", there is, e. g., no article on *Naturschutz*. The topic was not yet relevant in the years before 1889, when Matthias von Lexer worked on the N-articles. It developed to a complex discourse topic during the 20th century, an article *Naturschutz* is now available via zdl.org. This example shows that digital systems allow for a more timely reaction if it becomes apparent that a key term of a public discussion is not yet available. Digital systems also allow for interactive article management if the project teams want to implement such an option.

Apart from cases where on does not find a word in question, there are many examples for words and senses pertaining to social, cultural and discourse aspects that *are* explained and documented in different dictionaries or digital lexical systems. If, however, one looks for specific lexical fields, it is difficult to determine, whether the words and senses related to specific socio-cultural fields or aspects are documented in a systematic way. The work on the structure and the development of the Covid vocabulary at both the IDS and the DWDS has shown what it takes to document one vocabulary zone systematically over two years. So far, we have no publicly available instruments to check, if the words related to HIV/AIDS and many other topics are covered in a systematic way in our dictionaries and lexical systems.

There is another aspect where traditional lexicographical practices and the demands of describing and documenting social, cultural and discourse-related vocabulary are in conflict. Traditionally, word formations, which may be understood compositionally, were often not documented in dictionaries, even if they had a specific function and it would have been interesting, e. g., to document the time frame of usage in relation to topic careers. The expression *Dieseldesaster*, it seems, had a short career during the public discussion of VW's scandal about the manipulation of technical data. The term *Scud-Rakete* is another example for words that are related to specific fields of discourse. The career of military topics in public discourse is the basis of frequency profiles of words like *Scud-Rakete*, *Patriot-Rakete* or *Panzerhaubitze*. Lexical systems with a corpus component provide users with information about the use of a word, even if there is no fully-fledged article. In dwds.de, there is an article *Panzerhaubitze*,[8] both dwds.de and owid.de[9] have a number of corpus quotations mostly from newspapers.

[8] https://www.dwds.de/wb/Panzerhaubitze (last access: 19-06-2022).

[9] https://www.owid.de/artikel/224321 (last access: 19-06-2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

It is a task for future metalexicographical work to determine to what extent dictionaries of German cover different social, cultural or discourse-related vocabulary zones. One strategy to analyze the coverage of specific fields in dictionaries and lexical systems is the following: First, compile a body of texts which represent the field in question. Secondly, based on these texts make a list of words and/or senses that seem relevant. Thirdly, compare your list with what you already find in the dictionaries and lexical systems. Fourthly, decide and explain what is reasonably left out in the dictionaries (according to their or your criteria) and what is missing. To give a first example from military technology, a field with a very long history: In the 1980s the "Militärverlag der Deutschen Demokratischen Republik" published several booklets ("Militärtechnische Hefte") on subfields of military technology like "Kanonen und Haubitzen", "Panzerabwehr" or "Jagdflugzeuge". The whole series and its items mirror the knowledge system about military technology which in turn brings about a well-structured vocabulary with categories like (types of) military weapons (e. g. *Panzerabwehrlenkrakete*; *Granatwerfer*, *Panzerhaubitze*), types of persons (e. g. *Richtkanonier*), properties of military weapons (e. g. *Kaliber*, *Feuerkraft*) and many others. Such texts can be used to get an overview of the "specialized" side of knowledge fields and their vocabularies. To give a second example: between 1870 and 1930 there were many different reform movements with their respective textual cosmos and a specific vocabulary. A handbook on these movements (Kerbs/Reulecke 1998) combines a short portrait of the main protagonists and ideas with a short bibliography of important texts of each movement. One of these movements was vegetarianism. Since aspects of food and nutrition loom large in present day debates, a diagnosis of how the earlier stages of such discussions are documented in our dictionaries and digital lexical systems might take advantage of such a handbook. I will come back to the question of the coverage and accessibility of words and senses from a specific vocabulary zone in section 3.4.

3.2 Are the social, cultural, discourse-related aspects of the use of words part of their description and documentation?

There are many examples of German dictionaries containing articles on words with a specific social or cultural background. However, sometimes articles fail to explain and to make explicit this background. E. g., searching for *Lauberhüttenfest* and *Laubhüttenfest*, an expression referring to a religious holiday in the Jewish tradition, the DWb only provides two quotations and no descriptive text. FWb-Online refers to the article *laubhütte* ('aus belaubten Zweigen gebaute Hütte, zumeist zum Zwecke des Laubhüttenfestes errichtet')[10] where knowledge about the Laubhüttenfest and its cultural context is presupposed but not explained. The site woerterbuchnetz.de[11] provides access to an article in "Meyers Großes Konversationslexikon" (6th edition, 1905–1909) with encyclopedic information. The article *Laubhüttenfest* in dwds.de[12] gives an example how cultural aspects can be made explicit:[13]

[10] DeepL translates: *hut built from leafy branches, mostly erected for the purpose of the Feast of Tabernacles.*

[11] https://woerterbuchnetz.de.

[12] https://www.dwds.de/wb/Laubhüttenfest.

[13] DeepL translates: *Jewish religion -- seven-day festival celebrated by Jews in September or October to commemorate the Exodus of the Israelites from Egypt. -- Synonymous with Sukkot -- The festival is celebrated from the 15th to the 21st day of the first month of the Jewish lunar calendar* (*Tishri*), *five days after Yom Kippur, with the erection of a temporary leaf hut in which to eat and, if necessary, spend the night, and other customs.*

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Bedeutung DWDS-Vollartikel

jüdische Religion im September oder Oktober von den Juden begangenes siebentägiges Fest, das an den Auszug der Israeliten aus Ägypten erinnert **Das Fest wird vom 15. bis 21. Tag des ersten Monats des jüdischen Mondkalenders (Tischri), fünf Tage nach Jom Kippur, mit Errichtung einer provisorischen Laubhütte, in der gegessen und gegebenenfalls genächtigt wird, und weiteren Bräuchen begangen.**

Synonym zu **Sukkot**

KOLLOKATIONEN:

mit Adjektivattribut: das jüdische **Laubhüttenfest**

als Akkusativobjekt: das **Laubhüttenfest** feiern

als Genitivattribut: der Beginn des **Laubhüttenfests**

BEISPIELE:

So erinnert das **Laubhüttenfest**, das auf den 15. des siebten *[religiösen, bzw. weltlich ersten]* Monats festgelegt ist, an die Wüstenwanderung nach dem Auszug der Israeliten aus Ägypten und das Wohnen in Laubhütten[…]. [Badische Zeitung, 09.07.2013]

Bei öffentlichen Treffen mit Interessierten soll es […] um jüdisches Leben und jüdische Feste gehen – etwa um Pessach, das eng mit dem christlichen Osterfest verwoben ist, oder das **Laubhüttenfest**. [Saarbrücker Zeitung, 12.12.2020]

Im Judentum gab es zwei Erntefeste im Jahreskreislauf: das Pfingstfest als Getreide-Erntefest und das **Laubhüttenfest** als Wein- und Gesamt-Erntefest. [Mittelbayerische, 02.10.2020]

… 3 weitere Belege

What does it mean to describe and to document social, cultural or discourse-related aspects of the use of words and their senses in dictionaries or digital lexical systems? Which elements of articles may be used for these purposes? In the first place, the definitions and further descriptions of aspects of the usage provide an opportunity to make these relations explicit. Secondly, there are techniques of explicitly relating a word or a specific sense to a specific cultural, social or discourse-related "entity" or field of language use, e. g. by using a descriptor like "Jüdische Religion". Thirdly, it is essential, to give textual examples that show the relation to a specific cultural, social or discourse-related fields in the range of textual sources that are typical for the use of a given word or sense. This principle applies equally for the documentation of historical word usage. If most of the historical corpus quotations for a word like *Laubhüttenfest* come from anti-judaic sources, the corpus needs adjustment.

As for the aspect of documentation of word usage with textual examples, the question what texts are used for quotations is equally important. It seems to me that the strategy of using general, all-purpose corpora should be complemented by designing specific corpus components for specific social, cultural and discourse-related fields.

3.3 Do we get information about interrelations between words that belong to specific cultural, social, discourse-related vocabulary zones?

It is one thing to give information about social, cultural and/or discourse-related relations in a word article or in the article sections for specific senses of words. E. g., one can describe that a specific sense of Middle High German *trucken* ('humorally dry') is rooted in the system of humoral pathology and one can provide an explanation what the word means in old medical texts based on humoral thinking. It is a different task, however, to make clear in a dictionary or lexical system what the other words and senses are that belong to the same field. In printed dictionaries diasystematic predicates like "Jagdsprache" ('language of hunt-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

ing') or "Medizin" ('medicine') are not searchable. In digital systems such predicates could be searchable, but in some cases they are not, e.g. the descriptor "jüdische Religion" in the article *Laubhüttenfest* in dwds.de is not. Following the link "Synonym zu *Sukkot*", we find that the descriptor "jüdische Religion" is not yet part of the article *Sukkot*. But it is evident, that a kind of markup of articles or senses with searchable descriptors like "jüdische Religion", "Jagd", "Militärtechnik" and others is the way to go. This will allow to integrate social, cultural and discourse-related aspects in a system of faceted search in digital lexical systems.

Extending the perspective from the question "What are words related to X in a specific vocabulary zone?", the next step is the question for these vocabulary zones, their elements and the ways they may be addressed in dictionaries and digital lexical systems.

3.4 Are there ways of searching for or accessing specific vocabulary zones and the words that are their elements?

Dictionaries do not usually provide search facilities to access specific vocabulary zones and their elements. This includes those zones that are specifically related to social, cultural or discursive aspects. There are, however, notable exceptions.

Both Hermann Paul and Friedrich Kluge had the idea of providing an index to the words in their historical dictionaries, and the groups they built included social and cultural aspects, the notion of public discourses was not on their agenda. The latest edition of Paul's dictionary (2002) still includes the "Sachregister – Wegweiser zum Wortschatz" [Index – a guide to the vocabulary]. The criteria for grouping words together are manyfold, from types of linguistic development (e.g. "Bedeutungsgeschichte") to languages for specific purposes, prominent authors and many others. The connection with social and cultural aspects is evident in groups like "Aberglaube", "Amtssprache", "Anrede", "Begriffs- und Bedeutungsprägung" (with many entries on relevant key words for ideas and cultural items), "Bergmannssprache", "Bildungssprache", "Biologie", "Bühnensprache", "Computersprache", "DDR", "Demospruch/Losung", "Derbes, Obszönes" (which includes a small portion of the lexis of sexuality), to name but a few examples. The groups of this "Wegweiser" are organized alphabetically, therefore this index does not provide a structural system of the social and cultural world which is mirrored in the structure of the vocabulary.

This brings us to the historical thesaurus built into the OED. Its hierarchical, taxonomic organization is meant to make explicit a certain view of the external world, the mind and of society. It is obvious that a clear-cut division of such "realms" is problematic, because an aspect like "Health and disease" is not only a group in "The external world" but includes many expressions that refer to medical ideas or to social aspects of health care. Nevertheless, we have here an attempt to provide a complex guide to categories and different levels of subcategories down to groups like "educational buildings" with subgroups like "college or university buildings" and its 49 entries. In German, we have no such historical thesaurus as part of one of the historical dictionaries, nor do we have such built-in tools in the dictionaries for modern German. What we do have though are onomasiological dictionaries like the one by Franz Dornseiff, but they are not up to date and they are not interoperable given their idiosyncratic structures. There have also been discussions about the use of GermaNet as a part of WordNet, but the nodes in WordNet are not social, cultural or discourse-driven points of reference but terms that constitute so-called synsets. Nevertheless, this attempt to

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

organize lexicological substance in a non-alphabetical way deserves our attention with respect to the possibilities of implementing network structures beyond the alphabetic organization in digital lexicographic systems.

In addition, several dictionaries focus on specific social, cultural or discourse-related aspects of word usage. I already mentioned the discourse dictionaries produced by Heidrun Kämper and her teams that are combined with monographic investigations. In these books and in the web dictionaries a specific discourse topic is the organizing framework for the choice of articles and for their organization, e. g. in respect of the "Schulddiskurs" in post-WWII Germany. A completely different example is Ernest Bornemanns "Sex im Volksmund" (1984), an attempt to document the sexual vocabulary in German in an alphabetically organized dictionary and to combine the dictionary with a thesaurus. The problem with this endeavor is that it was not based on documented examples of "real" language use. In a "funny" passage, Bornemann tells us the story of his project. At some point, all kinds of people suggested sexual words and phrases to him, including sexworkers, which are introduced in the dedication with names like *Fischbüchsenpaula* or *Kitzler-Witzler*.

> Als die Barriere des Schweigens nach ein paar Jahren dann langsam abgetragen wurde, war der Informationsstrom allerdings kaum aufzuhalten. Am Telefon, an der Haustür, zu den erstaunlichsten Tages- und Nachtzeiten meldeten sich die erstaunlichsten Wesen, die meine Verwandten, Freunde und Nachbarn je erblickt hatten. (Bornemann 1984, unpag. Dedication)[14]

The construction of his "sexual thesaurus" is a good suggestion, if somewhat outdated in respect of the plurality of practices and new ways of thinking. More problematic is that there is no documentation for the lexical items from real word usage and that the material of the thesaurus is much broader than the items in the dictionary.

There are many more examples of publications that relate to specific aspects of the connection of word usage and cultural, social and discourse-related aspects in the history of German lexicography and lexicology, e. g. dictionaries for professional fields (e. g. Schirmer 1911; Kluge 1911) or monographs on economic sectors like forestry (e. g. Kehr 1964). Moreover, there are meta-lexicographical statements that emphasize the importance of the cultural and culture-pedagogical mission of German lexicography, most notably in the work of Oskar Reichmann (e. g. 2012).

4. Social, cultural and discourse-related aspects of word usage in next-generation lexical systems

As for the question: Can dictionaries (alone) with their alphabetical organization and with their "isolated" word-related articles alone provide cultural context and explain the complex lexical connections within cultural, social and discourse-related fields of word usage, it has become evident that it is possible to describe social, cultural and discourse-related aspects of word usage in current dictionaries and lexical systems of German, even if these options are not always used. Moreover, there are limitations that come from the alphabetical organization and word articles as the basic elements of dictionaries and lexical systems. Still, there

[14] DeepL translates: *When the barrier of silence was slowly removed after a few years, the flow of information could hardly be stopped. On the telephone, at the front door, at the most amazing times of the day and night, the most amazing beings ever seen by my relatives, friends and neighbors came forward.*

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

are a number of possibilities to further improve digital lexical systems in this respect. I propose the following suggestions.

1) Describe word usages in ways that explicitly include the social, cultural and/or discourse-related background of specific "senses" of words. As an option, this is already available, but it should become a *general principle* in future systems to use these options in one way or another.
2) If available, give links to encyclopedic information that describes cultural fields in which word usages are embedded, in a more "holistic" way. E. g., in order to explain the Medieval and Early Modern senses of *trucken*, *kalt* or *feucht* in old medical texts in the tradition of humoral pathology, it will not be possible to explain the whole system in each article. What can be done in dictionary articles is to briefly explain the role of the relevant sense in a conceptual system, e. g. humoral pathology, and to point the readers to information, where this system is characterized in a coherent way (cf. Gloning 2005). The question of how individual words and senses can be related to their cultural "surroundings" in digital systems is a major concern.
3) In order to overcome the shortcomings of the principle of isolated articles one can enrich or combine the dictionary with contributions that focus on specific aspects of word usage across individual articles. These can consist of specific "glossaries" to current topics like the Covid developments that have been produced both at the IDS and the DWDS. Blog posts can be used to answer questions like "What are old and new expressions for professions in German and where do we find them?" Earlier I have suggested so-called "Wortschatz-Miniaturen", small lexicological portraits that explain specific aspects of the structure and the history of vocabulary zones and that provide links to relevant dictionary entries (cf. Fritz 2020, p. 84 and chap. 2.12). This idea can be combined with thesaurus principles or with work that has been done in historical lexicology (e. g. Gloning 2003). The combination of discourse-related dictionaries and monographic investigations can be used for social and cultural aspects as well, e. g. in dissertation projects.[15]
4) A powerful system to overcome the limitations of alphabetic order is to use lexicological descriptors for different aspects of vocabulary organization for each sense of a word in a faceted search framework. E. g., if an entry like *Influencerin* (with only one sense) has descriptors like "expression for a profession", "expression for a female person", "english origin", one can formulate queries like: "Show all entries that fit the criteria: 'expression for a profession' and 'expression for a female person' and 'quotations from 1950–2022'", which would produce results like *Influencerin*, *Putzfrau*, *Managerin*, *Dramaturgin* and many others. This technique is much more flexible than the use of thesauri or ontologies, because it allows to combine criteria in different sets.[16]
5) Current work on specific developments like the Covid vocabulary has shown that it is fruitful to complement the criterion of frequency with the aim to document specific vocabulary zones. This should become a strategy of lexicographical work not only for

[15] Two Gießen projects may serve as examples, the work of Anna Pfundt on word usage in debates about women's suffrage around 1900 and Andre Pietsch's project on word usage in early texts on film and cinema.

[16] This is not to say that ontologies are not useful: E. g. the huge Krünitz encyclopedia from the 18th century has been enriched in woerterbuchnetz.de by markers that refer to the Dewey classfication which is extremely helpful. – For an example of extended search facilities beyond the alphabet see the IDS portal on loan words from German in other languages at: http://lwp.ids-mannheim.de/search/meta (last access: 19-06-2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

extraordinary situations but also for important social, cultural and discourse-related vocabulary zones and the textual cosmos in which the topic in question is situated.

5. To conclude

Words and their usages are in many cases closely related to or embedded in social, cultural, technical, discourse-related, ideological etc. contexts. This does not only apply to individual words and specific senses, but to many vocabulary zones as well. Moreover, the development of words is often related to aspects of socio-cultural evolution in a broad sense. In this paper, I have tried to first elucidate these kinds of connections between word usage and the socio-cultural. I have then tried to show, how these aspects are treated in traditional dictionaries and in digital lexical systems both in respect of words, senses and vocabulary zones. Finally, I have made suggestions as to how future digital systems might be improved and enriched.

References

DeepL = https://www.deepl.com/translator (last access: 19-06-2022).

Deutsch, A. (2022): Anforderungen an eine Bedeutungserklärung im Fachwörterbuch zu Zeiten von Google und Wikipedia – dargestellt am Beispiel des *Deutschen Rechtswörterbuchs*. In: Diehl, G./Harm, V. (eds.): Historische Lexikographie des Deutschen. Perspektiven eines Forschungsfeldes im digitalen Zeitalter. Berlin/Boston, pp. 37–56.

DWDS = https://www.dwds.de (last access: 31-05-2022).

Fritz, G. (2020): Darstellungsformen in der historischen Semantik. Gießen. [Online verfügbar: urn:nbn:de:hebis:26-opus-150846].

FWb-Online = https://fwb-online.de (last access: 31-05-2022).

Gloning, Th. (2003): Organisation und Entwicklung historischer Wortschätze. Lexikologische Konzeption und exemplarische Untersuchungen zum deutschen Wortschatz um 1600. Tübingen.

Gloning, Th. (2005): Wortbedeutung, Wortgebrauch, Wortschatzaufbau. Zu den Grundlagen und Aufgaben historischer Wörterbücher und historisch-lexikologischer Informationssysteme. In: Plate, R./Rapp, A. (eds.): Lexikographie und Grammatik des Mittelhochdeutschen. In Zusammenarbeit mit Johannes Fournier und Michael Trauth. Mainz/Stuttgart, pp. 61–97.

Gloning, Th. (2022): Historisches Vokabular des Jazz. Wort(schatz)geschichte in kommunikativen, kulturellen und digitalen Zusammenhängen. In: Diehl, G./Harm, V. (eds.): Historische Lexikographie des Deutschen. Perspektiven eines Forschungsfeldes im digitalen Zeitalter. Berlin/Boston, pp. 57–85.

Hacker, P. M. S. (2011): Language, language-games and forms of life. In: Padilla Gálvez, J./Gaffal, M. (eds.): Forms of life and language games. Frankfurt a. M. et al., pp. 17–36.

Kämper, H. (2012): Aspekte des Demokratiediskurse der späten 1960er Jahre. Konstellationen – Kontexte – Konzepte. Berlin/Boston.

Kämper, H.: Demokratiediskurs 1918–25. https://www.owid.de/wb/disk18/start.html (last access: 31-05-2022).

Kämper, H.: Schulddiskurs 1945–55. https://www.owid.de/wb/disk45/einleitung.html (last access: 31-05-2022).

Kämper, H.: Protestdiskurs 1967/68. https://www.owid.de/wb/disk68/start.html (last access: 31-05-2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Kehr, K. (1964): Die Fachsprache des Forstwesens im 18. Jahrhundert. Eine wort- und sachgeschichtliche Untersuchung zur Terminologie der deutschen Forstwirtschaft. Gießen.

Kerbs, D./Reulecke, J. (eds.) (1998): Handbuch der deutschen Reformbewegungen. Wuppertal.

Kluge, F. (1911): Seemannssprache. Wortgeschichtliches Handbuch deutscher Schifferausdrücke älterer und neuerer Zeit. Halle a. d. Saale.

OWID = Online-Wortschatz-Informationssystem Deutsch des Leibniz-Instituts für Deutsche Sprache. https://www.owid.de (last access: 31-05-2022).

Reichmann, O. (2012): Historische Lexikographie. Ideen, Verwirklichungen, Reflexionen am Beispiel des Deutschen, Niederländischen und Englischen. Berlin/Boston.

Schirmer, A. (1911): Wörterbuch der deutschen Kaufmannssprache auf geschichtlichen Grundlagen. Mit einer systematischen Einleitung. Straßburg.

Stötzel, G./Wengeler, M. (eds.) (1995): Kontroverse Begriffe. Berlin/New York.

Urban, H.-J./Ehlscheid, Ch. (2020): Generationengerechtigkeit. Grenzen und Potenziale eines sozialpolitischen Kernbegriffs. In: Aus Politik und Zeitgeschichte 70, 52–53, pp. 25–30. https://www.bpb.de/system/files/dokument_pdf/APuZ_2020-52-53_online_0.pdf.

Wittgenstein, Ludwig (1969): Tractatus logico-philosophicus. Tagebücher 1914–1916. Philosophische Untersuchungen. (= Schriften 1). Frankfurt a. M.

woerterbuchnetz.de = https://woerterbuchnetz.de (last access: 31-05-2022).

ZDL = Zentrum für digitale Lexikographie des Deutschen. https://zdl.org (last access: 31-05-2022).

Contact information

Thomas Gloning
Justus-Liebig-Universität Gießen
thomas.gloning@germanistik.uni-giessen.de

Acknowledgements

For their support and their comments I should like to thank Stefan Engelberg, Gerd Fritz, Dennis Kaltwasser, Annette Klosa-Kückelhaus and Christine Möhrs. – Your time and effort is much appreciated!

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Rufus H. Gouws

DICTIONARIES: BRIDGES, DYKES, SLUICE GATES

Abstract In a multilingual and multicultural society, dictionaries play an important role to enhance interlingual communication. A diversity of languages and different levels of dictionary culture demand innovative lexicographic approaches to establish a dictionary landscape that responds to the needs of the various speech communities. Focusing on the South African situation this paper discusses some aspects of a few dictionaries that contributed to an improvement of the local dictionary landscape. Using the metaphors of bridges, dykes and sluice gates it is shown how lexicographers need a balanced approach in their lemma selection and treatment. Whilst a too strong prescriptive approach can be to the detriment of the macrostructural selection, a lack of regulatory criteria could easily lead to a data overload. The lexicographer should strive to give a reflection of the actual language use and enable the users to retrieve the information that can satisfy their specific communication and cognitive needs. Such lexicographic products will enrich and improve the dictionary landscape.

Keywords Bilingualised dictionary; dictionary+; dictionary culture; dictionary portal; monolingualised dictionary; prescription

1. Introduction

Within the frame of the broad conference theme of lexicography in society, I have been asked by the organisers to discuss some aspects of dictionary landscapes in multilingual societies. As indicated in the title of my paper I am using the metaphors of dictionaries as bridges, dykes and sluice gates. I will apply these metaphors not only to the contents of dictionaries but also to a more comprehensive lexicographic process and to refer to some situations in a multilingual society that can have an effect on the planning, compilation, use and eventual success of dictionaries.

I am honoured to present this plenary paper as the AS Hornby lecture, and I gratefully acknowledge the massive contribution of AS Hornby to the field of lexicography. One of Hornby's major achievements, the monolingual learner's dictionary for Japanese students studying English, is proof thereof that not only a bilingual dictionary but also a monolingual dictionary can be a bridge between two languages. I will make reference to this approach in the paper.

2. Lexicography and society

The relation between lexicography and society can be complex and the relation holding within any given speech community seldom prevails in a similar way in other societies. Within a multilingual and multicultural environment, the dictionary landscape does not often reflect a balance between the different languages. Lexicographers compiling dictionaries for target users belonging to different speech communities need to negotiate the imbalances and complexities prevalent in the different languages, their speech communities and the available dictionaries and dictionary types. One of the major problems in any society and even more so in a multilingual society, is the lack of an established and comprehensive dictionary culture. Here I am not using the term *dictionary culture* in the way that Hausmann (1989, p. 13) used it to distinguish between user-friendliness in lexicography and a dictionary culture, with user-friendliness implying that lexicography adapts to society and

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

dictionary culture implying that society adapts to lexicography. I am using it in the way described by Gouws (2016, p. 111) to be a comprehensive umbrella term that includes the responsibility of both lexicography and society.

The lack of a dictionary culture or even of a rudimentary dictionary culture does not only impede lexicographers in their planning and eventual compilation of dictionaries, but it forms a dyke that separates these dictionaries from their intended target users. The dictionary landscape in any multilingual society is largely influenced by the nature and extent of a comprehensive dictionary culture or the absence thereof within each individual speech community.

In the remainder of this paper the focus will primarily be on the multilingual South African society with the emphasis on some dictionaries that add diversity to this dictionary landscape. The focus is not on the default general language dictionaries but rather on a few dictionaries that display innovative approaches to improve the nature of the dictionary landscape. Although the discussion is directed at the South African landscape, the different lexicographic endeavours could also be relevant to other multilingual societies.

Due to the reality of South Africa, the dictionary landscape shows both printed and online dictionaries. Online dictionaries are the default tools for certain user groups but for the majority of dictionary users and potential dictionary users, printed dictionaries currently still are the only lexicographic resources at their disposal. This situation poses some challenges to lexicographers and metalexicographers. In the transition from the printed to the online medium, the lexicographic practice led the way – with lexicographic theory following and having to play a catch-up game. Metalexicographers were slow in adapting theories that had originally been formulated for printed dictionaries to make provision for the emerging online dictionaries. Currently the metalexicographic discussion is dominated by the online medium. In South Africa lexicographic theory is also applied to ensure good online dictionaries. However, a real need remains for printed dictionaries and for an ongoing improvement of these dictionaries. Metalexicographers therefore need to formulate new models to enhance the quality of printed dictionaries and they need to embark on exciting endeavours to promote the transition to online dictionaries as well as the continued improvement of these lexicographic products.

3. Developing a dictionary culture

South Africa has eleven official national languages. Although there are huge differences in the size of the speech communities and the geographical distribution of the speakers of the eleven languages, these official languages are protected by the constitution. In practice they are not treated or used in an equal way. English dominates as lingua franca but also as language of the higher functions. Afrikaans, also due to support during the previous political era, is a fully standardised language that can be used at all levels of general and scientific communication. Due to, among others, the previous political landscape, the nine indigenous Bantu languages have not had the same support and do not show the same extent of development, especially in the domain of languages for special purposes. These differences between the languages are also evident in the dictionary landscape.

In principle, the future of the South African dictionary landscape should look positive. In addition to the lexicographic work of commercial publishers, the Pan South African Language Board, a government-funded organization, established to promote multilingualism,

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

to develop the eleven official languages, and to protect language rights in South Africa, founded a National Lexicography Unit (NLU) for each of the eleven official languages. The brief of these units is to develop the lexicographic landscape of their respective languages by compiling dictionaries for each speech community, with a comprehensive monolingual dictionary as the ultimate goal. When the NLUs were established in 2001 the playing field was not equal, nor was the dictionary culture of the different speech communities comparable. One model and even one dictionary type could not and still cannot be imposed on all the NLUs or on all the language groups in South Africa. This influenced the development of the different lexicographic projects and unfortunately today the dictionary landscape still shows vast differences between the different languages.

Wiegand (1998, p. 506) refers to a knowledgeable user ("ein kundiger Benutzer") and he identifies some features of such a user, but also of what he calls a non-knowledgeable user ("ein unkundiger Benutzer"). These features Wiegand identified include the familiarity, or lack thereof, of the user regarding the use of a dictionary – and the knowledge or non-knowledge such a user has of a specific dictionary. You can be a knowledgeable user of dictionary X but a non-knowledgeable user of dictionary Y. A knowledgeable user uses the dictionary in such a way that it conforms to the expectations of the compiler of the dictionary and the user has the proficiency and skills expected by the lexicographer. In contrast, the non-knowledgeable user, cf. Wiegand (1998, p. 507), does not have these skills that are presupposed by the lexicographer. These criteria of Wiegand confirm Hausmann's idea of a dictionary culture with society, the target users, having to adapt to lexicography – fulfilling the expectations of the lexicographer. The lack of sufficient knowledgeable users still prevents achieving an optimal dictionary landscape in South Africa.

However, when a dictionary culture is seen as a bidirectional process in which both society and lexicography play a significant role, one should not only work with the distinction between knowledgeable and non-knowledgeable dictionary users but also knowledgeable and non-knowledgeable lexicographers. Knowledgeable lexicographers have the skills and proficiency to plan and compile dictionaries that respond to the expectations, the lexicographic needs, and the reference skills of the target user. These skills and this knowledge needed by a lexicographer will not necessarily be the same when working in a monolingual compared to a multilingual society. In lexicographic research a lot of attention had been given to user studies. Lexicographer studies have not attracted enough attention. To what extent are the lexicographers in a multilingual and multicultural environment able to respond to the real lexicographic needs of diverse user groups – also within a single language? The dictionary landscape is not only determined by the available dictionaries but also by the dictionary culture and by the dictionary users and lexicographers who are primary participants in establishing the landscape.

A comprehensive dictionary culture demands that both lexicography and society need to adapt so that better dictionaries can be compiled and be used in an optimal way. This could help to ensure a better dictionary landscape.

4. Dictionaries: bridges, dykes and sluice gates

4.1 Bridges

The title of Bathe's book *Ianua Linguarum* (1615) – the gate of tongues – illuminates an important assignment of any dictionary – it should give access to data. Lexicographers

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

should be instrumental in making these data available to the target users and these users need to be proficient to execute a successful dictionary consultation by retrieving the required information from the data on offer.

Zgusta (1970, p. 294) already stated that "The basic purpose of a bilingual dictionary is to coordinate with the lexical units of one language those lexical units of another language which are equivalent in their lexical meaning". He also emphasised that the "fundamental difficulty of such a co-ordination of lexical units is caused by the anisomorphism of languages […]". Bilingual dictionaries are typical bridges in a multilingual society and this coordination is the typical bridging function of such a dictionary with linguistic, cultural and pragmatic features coming into play. When compiling bilingual dictionaries, lexicographers face challenges. It is not always possible to find exact equivalents to present in any bilingual dictionary. Lexicographers will be confronted with lexical gaps, and they need to counter them in the best possible way. The Nguni word *ubuntu* conveys a very specific cultural value and it does not have a direct equivalent in English. The word has to be included as lemma in a Zulu or Xhosa dictionary and the lexicographer could give a brief paraphrase of meaning like "good moral nature and human kindness". Knowledge of both the linguistic and the cultural aspects of these languages is of paramount importance to the lexicographer when coordinating their lexical units.

It is important that a lexicographer, especially in a multilingual environment, should adopt a comparative approach that takes cognizance of users from different speech communities. This could have an influence on the structure of, especially, the bilingual dictionary he/she is compiling. Responding to the question: "What do I want my user to be able to do with this dictionary?", a lexicographer might realise that the envisaged article structure of a dictionary might not accommodate all the data that should be included to support the target users. Lexicographers should be aware of the freedom they have to deviate from homogeneous article structures by employing clearly defined heterogeneous article structures. All articles will present at least an obligatory microstructure, but some articles may also present an extended obligatory microstructure that includes some items not relevant to all articles, e. g., a cultural note or footnote. In addition, within the frame structure of a printed dictionary an innovative variety of outer texts can be employed to increase the data distribution options. In an online dictionary, outer features, cf. Klosa/Gouws (2015), can be introduced and the lexicographer may even employ a data-pulling structure, cf. Gouws (2018), to enable access to dictionary-external sources.

Although bilingual dictionaries are the primary bridges in multilingual societies one should never underestimate the bridging value of monolingual dictionaries – provided, that they have been planned and compiled for a very specific situation of use. In this regard lexicographers can take guidance from the work of AS Hornby, and more specifically his *Idiomatic and Syntactic English Dictionary* (1942), later to be published internationally as *A Learner's Dictionary of Current English* (1948) and still later as *The Advanced Learner's Dictionary of Current English* (1952), cf. Cowie (1998). It is important to note that this first learner's dictionary was a monolingual product, and it is furthermore important to be aware of the environment in and for which this dictionary was prepared. In 1931 Hornby was invited by the linguist H. E. Palmer to join him in his work directed at vocabulary research at the Tokyo Institute for Research into English Teaching. According to Cowie (1998) this was almost ten years after Palmer had been commissioned to prepare a controlled vocabulary for Japanese middle schools. Palmer had already indicated the need for a special dictionary for the learner and the idea of a monolingual general-purpose dictionary designed particularly for ad-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

vanced Japanese learners of English had already been a topic of discussion. Hornby's knowledge of the needs of language learners in the Japanese situation guided the work towards the *Idiomatic and Syntactic English Dictionary* in 1942 and later the *Advanced Learner's Dictionary.* In the preface Hornby indicated that the dictionary had been compiled to meet the needs of foreign students of English and although not explicitly stated in the title of this dictionary a major feature of the *Idiomatic and Syntactic English Dictionary* was its clearly defined target user, i. e., the Japanese learner of English.

Although the user indicated by the term *learner* in Hornby's early learner's dictionary could be clearly identified as a Japanese learner of English as a foreign language this approach of working with such a well-defined target user no longer prevails in the modern-day lexicographic practice of monolingual dictionaries. Monolingual learner's dictionaries are typically compiled for learners of the specific language as a foreign language, but the native language of the target user is usually not specified. For a broad international market like that of the major English learner's dictionaries a general approach is in order because these dictionaries are not directed at target users from one specific language. However, in a multilingual country where dictionaries are bridges between members of the different speech communities more attention should be given to a more precise specification of the intended target users. Too often too little is known of the learners using these learner's dictionaries and this has definite implications for the success of this type of dictionary as a practical instrument.

By focusing specifically on the needs of Japanese learners of English AS Hornby could compile a dictionary that responded to the needs of the user of his dictionary but that could also negotiate their personal linguistic and cultural background. One would have expected that this approach would have been further developed by lexicographers of all monolingual learner's dictionaries. *Basiswoordeboek van Afrikaans* is a monolingual Afrikaans learner's dictionary compiled to help foreign language users learning Afrikaans. It was compiled for the South African market but fails to respond to specific problems that learners from some of the other South African languages will experience because not enough attention was given to the challenges faced by speakers of the Bantu languages who wanted to learn Afrikaans. The way in which a learner approaches a monolingual learner's dictionary is affected by the native language of the user and its traditions and cultures, cf. Atkins (1985:15). A dictionary that is too general cannot optimally suffice in a multilingual environment. Although it is commercially not viable to have a separate monolingual dictionary of, say Afrikaans, for each of the other South African languages, a single monolingual dictionary can present a generic approach complemented in either the articles or the outer texts by data directed at specific other languages, cf. Gouws (2015). In an online dictionary this can be achieved more easily.

When deciding on the way in which the native language of a user should play a role in the lexicographic presentation and treatment of a monolingual learner's dictionary the lexicographer needs to negotiate a variety of issues. These are issues regarding the structure of the language, the relation between the target language and that of the user, the culture of the speakers of the target language, the culture of the speakers of the native language, similarities and differences between the two languages, etc. In a multilingual and multicultural environment these considerations are even more compelling.

Bilingual dictionaries have a high usage frequency in multilingual societies and as practical instruments they play a significant role in the promotion of interlingual communication.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

These dictionaries should not only provide linguistic assistance but should also enhance a mutual understanding of different cultures. In this regard, the South African dictionary landscape has recently been enriched with excellent bilingual dictionaries, especially school dictionaries, with English as one of the treated languages. OUP South Africa has led the way with the publication of dictionaries with Northern Sotho, Zulu and Afrikaans respectively as the second member of the language pair. Pharos publishers has also contributed with, especially, their Afrikaans-English school dictionary. By enriching the dictionary landscape with school dictionaries, the foundation is laid for a process of life-long dictionary use. The introduction of good school dictionaries in South Africa also helps to avoid future lexicographic lost generations.

The bridging contribution of lexicography is not restricted to traditional bilingual dictionaries. Wiegand (2013, p. 285) refers to printed utility tools with formal properties of lexicographic nature, and this is also seen in the South African lexicographic landscape. Innovative endeavours, e. g., where the lexicographic work is complemented in a single source with other forms of language material result in a product that can be regarded as a dictionary+. Multilingual lexicographic products are not only bridges between the official languages of South Africa but are also employed to promote minority languages. One such example is found in N|uu, one of the few surviving non-Bantu click languages in Southern Africa and one of the most endangered languages on the continent.

4.1.1 A dictionary+

Efforts are currently made by a few of its remaining speakers to teach N|uu to descendants of the original speech community. Lexicography comes to the help again – an illustrated trilingual N|uu-Afrikaans-English reader: *Ouma Geelmeid ke kx'u ||xa||xa N|uu/Ouma Geelmeid gee N|uu.* (= Granny Geelmeid teaches N|uu) (Shah/Brenzinger 2016). This reader is divided into chapters in which words and expressions from a number of different thematic fields are presented, along with a few illustrations. In these thematic sections a variety of expressions are given in N|uu with translations into Afrikaans and English. In addition to the expressions illustrating the typical use of the language some chapters also contain single words from that semantic field with an illustration for each word. According to the authors "The contents of the reader and also the format are tailored towards the community needs in the N|uu teaching and learning efforts" (ibid., p. 10). By giving the expressions the reader adheres to a text production and translation function whereas the pictures satisfy a text reception and cognitive function. The lexicographic component is explicitly realised in two glossaries, N|uu-Afrikaans-English and Afrikaans-N|uu-English, presented as the final texts in this carrier of text types. These glossaries are preceded by illustrated charts of the various clicks, consonants and vowels of N|uu.

This reader is not a dictionary in the traditional sense of the word, but it contains lexicographic components complemented by other texts that present lexical, phonetic, orthographic and syntactic documentation of this endangered language. The reference in Wiegand (2013, p. 285) to printed utility tools with formal properties of lexicographic nature also applies to this dictionary. The principles of language documentation typically found in lexicographic work dominate this publication and the application of established lexicographic principles resulted in an innovative source of language documentation. The significance of this publication becomes clear when one is familiar with the linguistic situation in South Africa and the need to protect the endangered language of a part of society of which most of the members are non-literate. The target users of this readers are descendants of the N|uu

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

speech community. The genuine purpose of this reader is to “help students to learn to read and write N|uu, and even more importantly, to speak the language” (Shah/Brenzinger 2016, p. 10). As can be seen in figure 1 and 2 from the central list N|uu is the source language with Afrikaans and English as languages in which equivalents and translations are given.

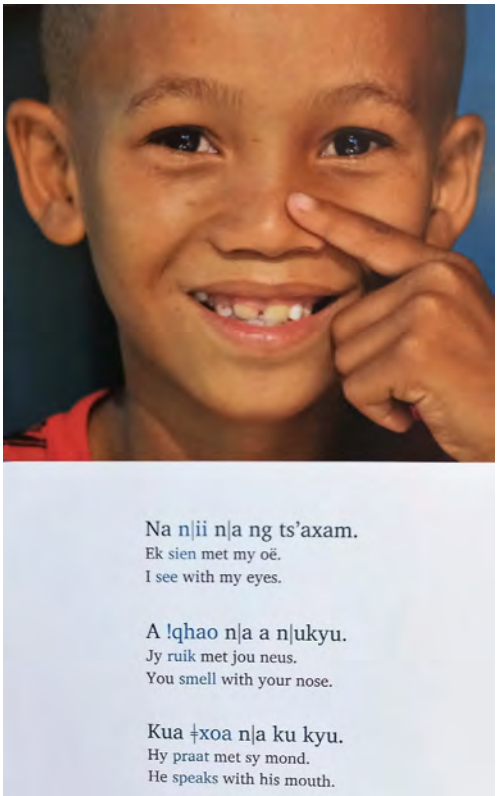


Fig. 1: from N|UU

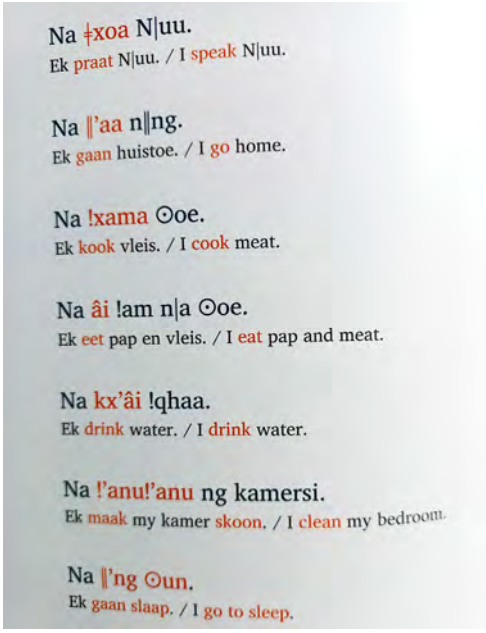


Fig. 2: from N|UU

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

The glossary in the back matter section contains the real lexicographic texts, i. e., two word lists: N|uu-Afrikaans-English (see fig. 3) and Afrikaans-N|uu-English.

	!qoarasi	!qoara	kortbeenboesmangras (deur skape gevreet; droë kort gras word gebruik om vuur mee te maak)	small grass (eaten by sheep; used to make fire when dry)
	!qoeke		om hande te klap	to clap hands
	!qunsi		motreën	drizzle
	!qui	!quike	as	ash
!QH	!qhaa		water	water
	!qhaa		langs (iemand, iets)	be next to (someone, something)
	!qhaaxu	ka !qhaaxuke	plaas	farm
	!qhâisi	ka !qhâi	spoor van diere of voetspoor van mense	footprint, spoor, track (e.g. of animals and people)
	!qhao		om (iets) te ruik	to smell (something)
	!qhobasi	!qhobake	hoodia (plant)	hoodia (plant)
	!qhoeke	ka !qhoeke	leeu	lion
	!qhûia		om vet te wees	be fat
!X'	!x'am		dagga	marijuana
	!x'aru	ka !x'aru	jagluiperd	cheetah
	!x'oa	ka !x'oa	volstruiskuiken	ostrich chick
	!x'uuke	ka !x'uu	voet	foot
ǂ	ǂama		bruin	brown

Fig. 3: from N|UU

The selection and ordering of the second and third languages in this reader and of the source language in the glossaries is not randomly done. Afrikaans is the first language of most of the target users and for them the N|uu words and expressions are readily accessible via Afrikaans. This dictionary offers a bridge from the known (Afrikaans) to the unknown (N|uu) and a basic treatment of the N|uu items. The more advanced user can eventually use the main access structure as constituted by the access route of the N|uu source language items. Given the multilingual environment the users are also presented with the relevant English equivalents. Within a specific linguistic landscape this dictionary responds to the specific multilingual communication and cognitive situation of its intended target user.

The structure and contents of this dictionary look quite simple, but this simplicity results from the execution of a well-devised plan to promote language use as well as the coordination of an endangered language and two official languages. In addition, the dictionary landscape is expanded. Such a lexicographic approach is important in a multilingual society.

4.1.2 Bilingualised dictionaries

Bilingualised dictionaries, cf. Nakamoto (1995), Laufer/Lindor (1997), also play an important bridging role in the South African dictionary landscape. Enhancing interlingual communication is not only done within a single dictionary but also by means of a series of dictionaries functioning as an interactive dictionary portal. Maskew Miller Longman published a series of foundation phase dictionaries (in the South African school system "foundation

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

phase" refers to the first three formal school years) that includes dictionaries for Afrikaans, Northern Sotho, Tswana, Xhosa and Zulu. These are monolingual dictionaries with a bilingual dimension. They are compiled for mother-tongue speakers of the specific language, but each dictionary article also contains an English translation equivalent as well as an English translation of the example sentence given to support the paraphrase of meaning.

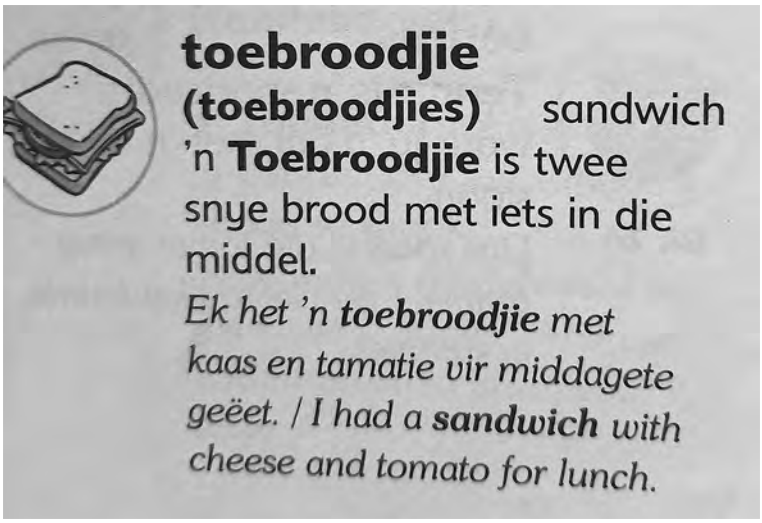


Fig. 4: from Grondslagfasewoordeboek

The back matter section of each dictionary in this series contains two alphabetical word lists. The first list includes all the words entered as lemmata in the central list with their English equivalents and a page number or page numbers where the source language word is treated. The second word list has English equivalents from the central list as source language items with the lemma from the primary language of the dictionary as equivalent, along with the page number or numbers where the item from the primary language is treated.

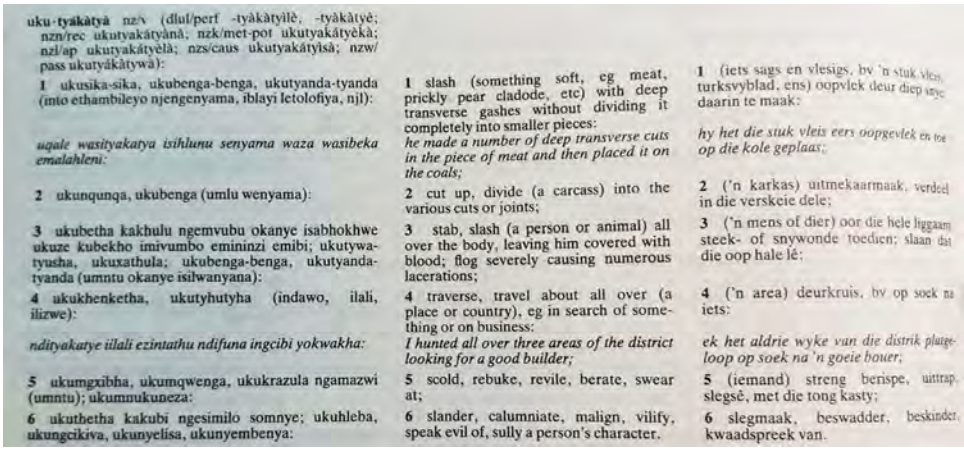
Important in a multilingual society is that each dictionary in this series is poly-accessible – either via the central list or via the back matter texts with their alphabetically ordered word lists presenting the two languages of the dictionary. Although these dictionaries are primarily monolingual – the paraphrase of meaning is only given in the source language of the central list – they can also be regarded as bilingualised dictionaries due to the presence of the English translation equivalents, example sentences and back matter word lists. As an independent publication each dictionary plays an important role in promoting the source language in combination with English as the lingua franca. In addition, and in response to the specific society, the dictionary series promotes multilingualism. To enhance interlingual communication all the dictionaries in this series show a comparable lemma selection. The lexical items presented in one of the monolingual English dictionaries of the publisher was used as basis for the macrostructural selection of all the dictionaries. These English words had been translated into the different languages and these equivalents were entered as lemmata in the respective dictionaries. Due to cultural and linguistic reasons some minor adaptations were made in the different dictionaries but to a large extent they display a comparable lemma selection. Consequently, the bridging does not only prevail between English and each one of the other languages individually. A user can move from the primary language of anyone of these dictionaries with English as bridging language to any of the other lan-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

guages. To illustrate this: the Xhosa dictionary offers English equivalents by means of which a Xhosa user can move from Xhosa to the English equivalent and then to the back matter text English-Tswana in the Tswana dictionary to finally reach the Tswana word that is an equivalent of the Xhosa word with which the search commenced. The comprehensive data distribution structure with the dictionary portal as a search domain and each individual dictionary as a search region, cf. Gouws (2021, p. 6), allows a retrieval of information from all the languages of the series and enhances the communicative potential of the South African society. This is a way of expanding the dictionary landscape by increasing the number of dictionaries available but also by elevating the communication potential in the specific multilingual society. Once these dictionaries are made available in online format the interlingual linking will be almost effortless.

4.1.3 Monolingualised dictionaries

Within a multilingual environment bilingualised dictionaries or monolingual dictionaries with a bilingual dimension can be complemented by monolingualised dictionaries or bilingual dictionaries with a monolingual dimension. In a linguistically and culturally diverse society like South Africa it is important to have dictionaries that can account for the lexicographic needs of the members of each speech community but can also guide the primary target users to other languages and can provide secondary users, i. e., users from one or more different South African languages, access to the primary language of the dictionary. A dictionary that achieves exactly this purpose is the *Greater Dictionary of Xhosa*. This three-volume dictionary can be regarded as a trilingual dictionary with a strong monolingual dimension – in the sense that the treatment has been enhanced through the inclusion of items usually only associated with monolingual dictionaries. Each page displays partial article stretches spread over three columns, with columns for English and Afrikaans running parallel to that of the Xhosa column.



uku-tyàkàtyà nz/v (dlul/perf -tyàkàtyìlè; -tyàkàtyè; nzn/rec ukutyakàtyànà; nzk/met-pot ukutyakàtyèkà; nzl/ap ukutyakàtyèlà; nzs/caus ukutyakàtyìsà; nzw/pass ukutyàkàtywà):

1 ukusika-sika, ukubenga-benga, ukutyanda-tyanda (into ethambileyo njengenyama, iblayi letolofiya, njl):

ugale wasityakatya isihlunu senyama waza wasibeka emalahleni:

2 ukunqunqa, ukubenga (umlu wenyama):

3 ukubetha kakhulu ngemvubu okanye isabhokhwe ukuze kubekho imivumbo emininzi emibi; ukutywatyusha, ukuxathula; ukubenga-benga, ukutyandatyanda (umntu okanye isilwanyana):

4 ukukhenketha, ukutyhutyha (indawo, ilali, ilizwe):

ndityakatye iilali ezintathu ndifuna ingcibi yokwakha:

5 ukumgxibha, ukumqwenga, ukukrazula ngamazwi (umntu); ukumnukuneza:

6 ukuthetha kakubi ngesimilo somnye; ukuhleba, ukungcikiva, ukunyelisa, ukunyembenya:

1 slash (something soft, eg meat, prickly pear cladode, etc) with deep transverse gashes without dividing it completely into smaller pieces:
he made a number of deep transverse cuts in the piece of meat and then placed it on the coals;

2 cut up, divide (a carcass) into the various cuts or joints;

3 stab, slash (a person or animal) all over the body, leaving him covered with blood; flog severely causing numerous lacerations;

4 traverse, travel about all over (a place or country), eg in search of something or on business:
I hunted all over three areas of the district looking for a good builder;

5 scold, rebuke, revile, berate, swear at;

6 slander, calumniate, malign, vilify, speak evil of, sully a person's character.

1 (iets sags en vlesigs, bv 'n stuk vleis, turksvyblad, ens) oopvlek deur diep sny daarin te maak:

hy het die stuk vleis eers oopgevlek en toe op die kole geplaas;

2 ('n karkas) uitmekaarmaak, verdeel in die verskeie dele;

3 ('n mens of dier) oor die hele liggaam steek- of snywonde toedien; slaan dat die oop hale lê;

4 ('n area) deurkruis, bv op soek na iets:

ek het aldrie wyke van die distrik pluisloop op soek na 'n goeie bouer;

5 (iemand) streng berispe, uittrap, slegsê, met die tong kasty;

6 slegmaak, beswadder, beskinder, kwaadspreek van.

Fig. 5: from the Greater Dictionary of Xhosa

This article structure resembles what Wiegand/Feinauer/Gouws (2013, p. 328) call a block article. It differs, however, because each block is not an article but only a partial article because only the Xhosa block has a lemma sign. It can be regarded as a blocked article consisting of three partial blocks.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

The Xhosa column contains partial blocks that could function as fully-fledged articles in a monolingual dictionary. This partial block satisfies the minimum criteria of a basic article because it has its own comment on form and a comment on semantics. For each lemma the treatment in the first partial block is executed by means of Xhosa items, as could be expected in a monolingual dictionary. The second and third columns contain partial blocks presenting partial articles that consist only of a comment on semantics containing the respective English and Afrikaans equivalents or translations of the Xhosa paraphrases of meaning as well as example sentences in articles where the Xhosa section has example sentences. The outer access structure of the central list of this dictionary has a single search route that guides a user to the Xhosa lemma sign. The search route of the inner access structure guides a user to the items in the Xhosa search zones and then to the subsequent horizontally ordered English and Afrikaans partial articles.

The specific article structure of this dictionary is not for metalexicographic cosmetic reasons, but it is motivated by the relation between lexicography and society. In the preface to this dictionary the editors say:

> The three languages used side by side bring to mind the eventful history of interaction, co-operation and conflict, and the ferment days now past. However, the Dictionary is making its appearance at a time when the peoples of Southern Africae learning the need for greater understanding and acceptance of one another, and it is hoped that the use of these volumes will in some way contribute to this process. (Pahl 1989, p. viii)

Within a multilingual and multicultural society, the dictionary has a primary target user group, but it equips these users with more than a mere knowledge of their own language. It enhances interlingual communication.

4.2 Dykes

The question that should dominate all decisions regarding the contents of a dictionary, i. e. "What do I want my user to be able to do with the dictionary?" should also determine whether a lexicographer adopts a prescriptive, descriptive or proscriptive approach, cf. Bergenholtz (2003) and Bergenholtz/Gouws (2010), when it comes to the selection of items to be included in any given dictionary.

Dictionaries focusing on a presentation and treatment of the language for general purposes for a general target user group, not for school students, should avoid a dyke function that prohibits the inclusion of items that belong to the subject matter of the specific dictionary. These dykes could be of a linguistic, ideological, or cultural nature or could merely reflect the personal bias of the lexicographer.

In a multilingual society language contact is a normal phenomenon that occurs on a daily basis. In their reflection of the actual language usage lexicographers have to take cognizance of the results of this contact and, depending on the type of dictionary they compile and the genuine purpose of that dictionary, they have to plan the way in which their dictionaries should negotiate this. The dictionary landscape of a multilingual and multicultural country like South Africa should bear witness of the linguistic realities and the fact that no language in this society exists in isolation.

One can easily underestimate the extent of the influence of language contact with languages not only borrowing words from other languages but also lending words to other languag-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

es. Schoonheim (2021, p. 169) distinguishes between loanwords, i. e., those lexical items borrowed from other languages, and export words, i. e., those lexical items that are lent to other languages. Where there is a dominant language or lingua franca in a multilingual society, that language will often be the exporting language. In South Africa all the other ten official languages contain a variety of loanwords from English. However, dictionaries also show the extent to which South African English has not only exported to but has borrowed from other languages. *A Dictionary of South African English on Historical Principl*es (Silva 1996) gives ample proof of the way in which South African English has been influenced by the other South African languages. The lemma selection of this dictionary is restricted to borrowings from the other South African languages. From a linguistic perspective this dictionary acts as a bridge that displays the results of language contact with each borrowed form functioning as a miniscule communication bridge between English and one of the other languages.

Part of the bridging assignment of dictionaries is to include established loan forms to ensure the best possible interlingual comprehension. A too strong prescriptive approach, often motivated by misplaced linguistic purism or language nationalism, results in a dictionary becoming a dyke that isolates the dictionary from the surrounding language use – and from the speakers of that language. In the early decades of the previous century Afrikaans had to establish itself as a national language alongside the world language English. Although Afrikaans and English functioned together and a bidirectional influence existed, linguists and lexicographers tried to rid Afrikaans as far as possible from English influence. Employing a strong prescriptive approach many direct translations from English as well as English loan words were excluded from the dictionaries in spite of their occurrence in daily communication. In bilingual dictionaries with Afrikaans and English as language pair, see Bosman/Van der Merwe (1936) and Bosman/Van der Merwe/Hiemstra (1984), these anglicisms were replaced by Dutchisms and Germanisms – words and expressions that portrayed artificial and non-natural language use in Afrikaans. Typical Afrikaans words like *geboortemerk* (birth mark), *boekmerk* (bookmark), *rughand* (backhand) were excluded because they are direct loan translations from English. In their place the Dutch forms *moedervlek* and *boeklêer* and the unnatural form *handrug* were included. These substituting forms were not part of the active Afrikaans language use, and their inclusion diminished the representativeness of the dictionaries. Fortunately, things have changed. A more descriptive approach and an acknowledgement of the naturalness of language contact and the inevitable inclusion of loan forms and loan translations as well as the emergence of representative corpora helped to remove many dykes from the South African dictionary landscape.

Dykes are also created due to language-political issues, e. g., the standardisation process of a language with different dialects. A biased and one-sided standardisation process could form a dyke that prevents numerous forms from being considered for inclusion in a dictionary. This has also happened in the South African landscape. Mojela (2008, p. 119) discusses what he calls a "strict and narrow standardization" of Sesotho sa Leboa (Northern Sotho) that resulted in the exclusion of many dialectal forms and that imposed a standard language on the speech community that was foreign to many of them. As a result, some dialects were stigmatized and regarded as inferior. This dyke separating exclusion from inclusion often does not have an objective linguistic motivation. Consequently, Mojela (ibdi., p. 129) believes that lexicographers are faced with the challenge of bridging the gap between the standard language and those dialects that had been stigmatized. Here dictionaries should

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

not be dykes but rather bridges "in order to make the standard language acceptable to all the communities …" This should also guarantee the unity and stability of Sesotho sa Leboa.

One of the problems Mojela refers to is that some of the established corpora used by lexicographers did not include lexical items from the side-lined dialects. These corpora strengthened the dyke and supported the exclusion of words frequently used by speakers of the inferior dialects. This problem has been overcome in some of the more recent dictionaries, ensuring that their bridge function surpasses their dyke function.

4.3 Sluice gates

The metaphors of dictionaries as bridges, dykes and sluice gates do not only apply to the macrostructural coverage of a dictionary but can also be used with regard to other structures and procedures in the lexicographic process. Sluice gates can be interpreted in two ways: the opening of a sluice so that water can flow freely, or a type of lock in e. g., a river to manage the water flow and water level. Both these senses are relevant when using sluice gates as a metaphor in a discussion of dictionaries.

Looking at dictionaries as bridges, the enriching value of language contact has already been identified – as well as the unfortunate puristic attempts to create dykes to prevent this influence. Lexicographers need a well-balanced approach, guided by the reality of actual language use, to negotiate the functions of their dictionaries as bridges, dykes and sluice gates. Specific linguistic and lexicographic circumstances can also play a determining role, but a single dictionary can present all three these functions.

In the development of monolingual dictionaries in Afrikaans the comprehensive multivolume *Woordeboek van die Afrikaanse taal* (Dictionary of the Afrikaans language) (WAT), has played a significant role – and is still playing that role. This project was started in 1926 when there still was a lack of both other general monolingual Afrikaans dictionaries and Afrikaans special field dictionaries. The comprehensiveness of a comprehensive dictionary prevails on at least three levels: the lexical items included for treatment, the data types allocated to each article and the extent of the treatment. With regard to the lexical coverage and the extent of the treatment, the WAT opened the sluice gates. As one can expect from a dictionary belonging to this typological category, it contains a comprehensive selection of lexical items from the general language. In the absence of special field dictionaries many terms from a variety of subject fields that would not typically qualify for inclusion in a general language dictionary had been entered as lemmata. This created a lexical data overload because the dictionary contained items that should not have been lemma candidates for a general monolingual dictionary. Although there still was a lack of special field dictionaries, a general dictionary was not the venue where interested users would look for these items. This lexical overflow was detrimental to the focus and the genuine purpose of the WAT and impeded its progress. Changes in the dictionary landscape and the emergence of a range of other Afrikaans dictionaries convinced the editors of the WAT to adjust their lemma selection policy to close the sluice gates for some items.

Even in a comprehensive dictionary lexicographers must be aware of the slogan "less is more," although *less* does not always have the same value. Roughly during the period 1965–1985 the WAT, riding the wave of comprehensiveness, opened the sluice gates for certain types of data, especially data accommodated in the search zones for the paraphrases of meaning. An inflation of encyclopaedic data dominated these articles and impeded rapid

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

access to the core data in these search zones. Another type of sluice gate was needed: a type of lock to manage the data flow and data level. In the WAT the appropriate data level was found by a balance between a flow of lexicographic and non-lexicographic data and the regulating value of lexicographic theory. Following a lot of criticism from linguists and metalexicographers, the editors of the WAT devised a new data distribution plan for the dictionary articles with clearly defined criteria for the nature and extent of data provided in the paraphrase of meaning, cf. Botha (2003). This presentation of data bridges a knowledge gap and successfully assist users in retrieving the necessary information without stumbling over non-relevant data. In this regard the WAT has become an example for monolingual lexicographic work in the other South African languages.

In a multilingual country like South Africa that has English as a dominant language it is natural, predictable and acceptable to have English exporting words and expression to other languages. A balance is required because a random opening of the lexical sluice gates can result in languages being flooded by unnecessary loan words. Yet again, dictionaries have to reflect the actual language use, but they could also provide guidance and even issue a warning when needed. A mere transliteration of English words often results in an increase of the loan word stock of the indigenous South African languages. This is in spite of the fact that the lexicons of these languages often do have appropriate words available. The indigenous African languages often lack enough special field and technical terms, and loan words are accepted and welcomed. But not to replace existing words and terms. Here the sluice gates need to be closed so that these languages can develop and offer their speech communities the option of expressing themselves in all spheres of life in their mother language.

The Northern Sotho equivalent for the word *aeroplane* is *sefofane* (literally an object that flies). According to Makua (in preparation) some Northern Sotho speakers who are used to transliterating from English are using the form *folaematšhene* which is a borrowed term, a transliteration of *flying machine*. For a cell phone the transliteration *selefoune* has been used although Northern Sotho had already in the early years of mobile phones been enriched with its own word *sellathekeng* – "it cries/rings on the hips". As translation equivalent for *car* Northern Sotho has the word *sefatanaga* but the opened sluice gates allowed the transliteration *mmotoro*. According to Hlungwane (in preparation) there is a need for Northern Sotho (and other African language) dictionaries to provide their users with Northern Sotho items that are established forms in the language although they function alongside loan words and transliterations. The opening of the sluice gates should not endanger a language.

As authoritative sources dictionaries could show both the indigenous and the loan forms. Here lexicographers could adopt a proscriptive approach, cf. Bergenholtz (2003) and Bergenholtz/Gouws (2010). Such an approach could imply that a dictionary presents both these forms, but the lexicographers express a preference – which might be subjective or biased but could also be based on linguistic and cultural priorities as well as corpus evidence. The article structure may even allow the use of a text box or an article-internal footnote to motivate the specific preference.

Dictionaries need to contribute to the development of a language, and this can also be achieved by sluice gates that increase lexicotainment. When it comes to the inclusion of neologisms in dictionaries there are criteria determining when the usage frequency of a given form justifies its inclusion as lemma in a general language dictionary. Significant deviations from the traditional inclusion policies of neologisms were witnessed regarding

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

COVID-19 neologisms where an immediate lexicographic response was required, cf. some of the papers from the Globalex workshops on lexicography and neology (Klosa-Kückelhaus/Kernerman (in print)). In South Africa Afrikaans and the African languages need to expand their vocabularies. This is not only done by opening the sluice gates that allows borrowing from English but also by finding new words as non-borrowed translation equivalents for some English words. A couple of linguistic entrepreneurs in the educational environment proposed the idea of a dictionary with suggestions of new Afrikaans words for existing English forms. People were invited to submit their own neologisms and the *Wilde woordeboek* (Wild dictionary) (Van Niekerk/Basson/Grobler) entered the dictionary landscape. This dictionary was evidence of the innovative ideas of members of the Afrikaans speech community and showed the creative potential of the language and its contribution to the dictionary landscape. The *Wilde woordeboek* is a sluice gate that channelled linguistic creativity and enhanced the growth and development of Afrikaans.

5. Conclusion

The dictionary landscape in the multilingual and multicultural South Africa is diverse and the lexicographic standard of the different languages is not equal and does not display a parallel development. However, a variety of dictionary types and innovative lexicographic projects in different languages offer numerous interlingual bridging and collaboration opportunities. Dictionaries also have a dyke and a sluice gate function that plays a regulating role in the lexicographic presentation of linguistic forms.

A major problem is the lack of a comprehensive dictionary culture. To solve this problem joint ventures by lexicography and society are needed. The better the dictionary culture, the better the dictionary landscape and the less cumbersome the bridging between different languages.

References

Atkins, B. T. (1985): Monolingual and bilingual learners' dictionaries: A comparison. In: Ilson, R. (ed.): Dictionaries, lexicography and language learning. Oxford, pp. 15–24.

Bathe, W. (1615): Ianua linguarum. London.

Bergenholtz, H. (2003): User-oriented understanding of descriptive, proscriptive and prescriptive lexicography. In: Lexikos (13), pp. 65–80.

Bergenholtz, H./Gouws, R. H. (2010): A functional approach to the choice between descriptive, prescriptive and proscriptive lexicography. In: Lexikos (20), pp. 26–51.

Bosman, D. B./Van der Merwe, I. W./Hiemstra, L. W. (1984): Tweetalige woordeboek/Bilingual dictionary. Cape Town.

Bosman, D. B/Van der Merwe, I. W. (1936): Tweetalige woordeboek. Cape Town.

Botha, W. F. (2003): Die impak van die leksikografieteorie op die samestelling van die Woordeboek van die Afrikaanse Taal. PhD thesis. Stellenbosch University.

Botha, W. F. (ed.) (1951–): Woordeboek van die Afrikaanse Taal. Stellenbosch.

Cowie, A. (1998): A. S. Hornby: A centenary tribute. In: Fontenelle, T. (ed.) (1998): EURALEX 98 Proceedings. Liège, pp. 3–16.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Gouws, R. H. (2018): ʻn Leksikografiese datatrekkingstruktuur vir aanlyn woordeboeke In: Lexikos 28, pp. 177–195.

Gouws, R./Potgieter, L./Burgess, S. (eds.) (2010): Grondslagfasewoordeboek Afrikaans/English. Cape Town.

Gouws, R. H. (2015): Wie is die teikengebruikers van eentalige aanleerderwoordeboeke? In: Tydskrif vir Geesteswetenskappe (55/3), pp. 343–355.

Gouws, R. H. (2016): Op pad na ʻn omvattende woordeboekkultuur in die digitale era. In: Lexikos 26, pp. 103–123.

Gouws, R. H. (2021): Expanding the use of corpora in the lexicographic process of online dictionaries. In: Piosik, M. et al. (eds.): Korpora in der Lexikographie und Phraseologie. Berlin, pp. 1–20.

Gouws, R. H. et al. (eds.) (2013): Dictionaries. An international encyclopedia of lexicography. Supplementary volume: recent developments with focus on electronic and computational lexicography. Berlin.

Gouws, R. H./Feinauer, I/Ponelis, F. (eds.) (1994): Basiswoordeboek van Afrikaans. Pretoria.

Hausmann, F. J. (1989): Die gesellschaftlichen Aufgaben der Lexikographie in Geschichte und Gegenwart. In: Hausmann, F. J. et al. (eds.) (1989–1991): Wörterbücher. Dictionaries. Dictionnaires. An international encyclopedia of lexicography. Berlin, pp. 1–19.

Hlungwane, M. (in progress): A model for a bilingual online Northern Sotho linguistics special field dictionary. PhD thesis. Stellenbosch University.

Hornby, A. S./Gatenby, E. V./Wakefield, H. (1942): Idiomatic and syntactic English dictionary. Oxford. (Photographically reprinted and published as a learner's dictionary of current English by Oxford University Press, (1948); subsequently, in 1952, retitled The advanced learner's dictionary of current English. Tokyo).

Klosa, A./Gouws, R. H. (2015): Outer features in e-dictionaries. In: Lexicographica 31, pp. 142–172.

Klosa-Kückelhaus, A./Kernerman, I. (eds.) (in print): Lexicography and COVID-19 Neologisms. Berlin.

Laufer, B./Lindor, H. (1997): Assessing the effectiveness of monolingual, bilingual, and "bilingualised" dictionaries in the comprehension and production of new words. In: The Modern Language Journal 81/ii, pp. 189–196.

Makua, M. B. (in progress): The treatment of translation equivalence in bilingual dictionaries with English and Sepedi as language pair. PhD. thesis. Stellenbosch University.

Mojela, V. M. (2008): Standardization or stigmatization? Challenges confronting lexicography and terminography in Sesotho sa Leboa. In: Lexikos 18, pp. 119–230.

Nakamoto, K. (1995): Monolingual or bilingual? That is not the question: the "bilingualised" dictionary. In: Kernerman Dictionary News, January 1995, pp. 2–7.

Pahl, H. W. (ed.) (1989): The greater dictionary of Xhosa, Volume 3. Alice.

Schoonheim, T. (2021): Words crossing borders. In: Lexicographica 37, pp. 161–175.

Shah, S./Brenzinger, M. (eds.) (2016): Ouma Geelmeid ke kx'u ||xa||xa N|uu/Ouma Geelmeid gee N|uu. Cape Town.

Silva, P. (ed.) (1996): A dictionary of South African English on historical principles. Cape Town.

Van Niekerk, F./Basson, J./Grobler, K. (eds.) (1997): Wilde woordeboek. Pretoria.

Wiegand, H. E. (1998): Wörterbuchforschung. Berlin.

Wiegand, H. E. (2013): Gedruckte Gebrauchsgegenstände mit lexikographischen Formeigenschaften. In: Lexicographica 29, pp. 285–307.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Wiegand, H. E./Feinauer, I./Gouws, R. H. (2013): Types of dictionary articles in printed dictionaries. In: Gouws, R. H. et al. (eds.): Dictionaries. An international encyclopedia of lexicography. Supplementary Volume: Recent developments with focus on electronic and computational lexicography. Berlin, pp. 314–366.

Zgusta, L. (1971): Manual of Lexicography. The Hague.

Contact information

Rufus H. Gouws
Department of Afrikaans and Dutch
Stellenbosch University
rhg@sun.ac.za

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Nicola McLelland

WOMEN IN THE HISTORY OF LEXICOGRAPHY

An overview, and the case of German

Abstract This paper first attempts a state-of-the art overview of what is known about women in the history of lexicography up to the early twentieth century. It then focusses more closely on the German and German-English lexicographical traditions to 1900, examining them from three different perspectives (following Russell's 2018 study of women in English lexicography): women as users and dedicatees of dictionaries; women as contributors to and compilers of lexicographical works; and (in a very preliminary way) women and female sexuality as represented in German/English bilingual dictionaries of the eighteenth and early nineteenth centuries.

Russell (2018) was able to identify some 24 dictionaries invoking women as patrons, dedicatees or potential users before 1700, and some 150 works in English lexicography by women between 1500 and 1900, besides the contribution of hundreds of women as supporters and helpers, not least as unpaid readers and sub-editors for the *Oxford English Dictionary*. Equivalent research in other languages is lacking, but this paper presents some of the known examples of women as lexicographers. The evidence tends to support Russell's finding for English, that women were more likely to find a place in lexicography outside the mainstream: sometimes in a more private sphere (like Hester Piozzi); often in bilingual lexicography (such as Margrethe Thiele, working on a Danish-French dictionary), including missionary and or colonizing activity (such as Cinie Louw in Africa, Daisy Bates in Australia); and in dialect description (Coronedi Berti in Italy, Luisa Lacal and María Moliner in Spain).

Within the German-speaking context, women who participated in lexicographical work themselves are hard to identify before the late nineteenth century, though those few women who did have access to education were often engaged in language learning, including translation activity, and they were likely users of bilingual and multilingual dictionaries. Christian Ludwig's (1706) English-German dictionary – the first of its kind – was dedicated to the Electoral Princess Sophia of Hanover. Elizabeth Weir may have been the first named female compiler of a German dictionary, with her bilingual *New German Dictionary* (1888). Rather better known are the cases of Agathe Lasch and Luise Pusch, who, as pioneering women in the field of German linguistics, ultimately led major lexicographical projects documenting German regional varieties in the first half of the twentieth century (Middle Low German and Hamburgish in the case of Lasch; the Hessisch-Nassau dialect dictionary in the case of Berthold).

In the light of existing research on gender and sexuality in the history of English lexicography (e. g. Iamartino 2010; Turton 2019), I conclude with a preliminary exploration how woman and sexuality have been represented in dictionaries of German and English, taking the words *Hure* and *woman* in bilingual German-English dictionaries of the eighteenth and nineteenth centuries as my case studies.

Keywords Lexicography, German, women, Hester Piozzi, Margrethe Thiele, Cinie Louw, Theodor Arnold, Christian Ludwig, Elizabeth Weir

1. Introduction

Lindsay Rose Russell's ground-breaking study of women in English-language dictionary-making (Russell 2018) is important not just for the history of English lexicography, but also as a model for future work in other language traditions. Russell first unpicks the standard narrative, that in early English lexicography, women were, when invoked as potential dictionary users, 'useful as a passive and ignorant audience', an 'exploitable but ultimately expendable, uneducated demographic', and so one that ceased to be mentioned after about 1660 (Russell 2018, p. 30). James Murray, the first editor of the *Oxford English Dictionary* (*OED*), suggested in 1900 that the supposed 'elegant' ignorance of women, with their sys-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

tematically greatly reduced access to education, could mask the 'merely shameful ignorance of men' (Russell 2018, p. 34, citing Murray 1900, p. 32). Russell is able to show, in contrast, that

> invocations of women had been both genuine and genuinely successful in establishing a popular foothold for the genre [...] women's early involvements in dictionary making and use did not abate, but continued long after the seventeenth century' (Russell 2018, p. 34).

Russell (2018, pp. 41, 43) was able to uncover 24 dictionaries published between 1500 and 1700 which name or invoke women, whether as individual dedicatees, as individual inspiration (as former pupils, for example), or as a class of intended users. What is more, between 1500 and 1900, Russell finds some 150 lexicographical projects involving English undertaken by women, of which about a quarter are bi- or multilingual (Russell 1918, pp. 73, 76–105, Tables 3.1 and 3.2). Russell's definition of lexicographical activity is deliberative expansive, precisely because many women's activity lies at the margins of mainstream, archetypical dictionary-making, and is often linked to language learning, missionary activity, documenting local dialects, or focused on the domestic sphere in some way. An example of the latter is Mary Evelyn's (1690) *Mundus Muliebris: Or, The Ladies Dressing-Room Unlocked, Together with the Fop-Dictionary, Compiled for the Use of the Fair Sex* (Russell 2018, pp. 68–71).

Hester Lynch Piozzi (1740–1821) is emblematic of much of what Russell seeks to show about women and English lexicography. Piozzi was exceptionally well educated in several languages – 'till I was half a prodigy', in her own words – and she wrote and published work herself (Russell 2018, p. 143). She was a close friend and associate of Dr Samuel Johnson, compiler of the epoch-making *Dictionary of the English Language* (1755), and he relied heavily on Piozzi's collegiality and patronage: he had rooms in her house and used her library. Russell also makes a strong case that Piozzi's role in recording the history of the great man Johnson's work has been marginalized – her *Anecdotes of Dr Johnson* (1786), written in three months on her honeymoon after Johnson's death in 1784, was the first such account of Johnson's life, and is an important source. Yet it is, Russell suggests, backgrounded in Reddick's otherwise excellent (1990) account of Johnson's work on the dictionary, just as Piozzi's involvement in Johnson's life is backgrounded compared to the role played by Johnson's wife (Russell 2018, pp. 170 f.).

More than supporting Johnson, though, Piozzi was also a lexicographer in her own right. Her *British Synonymy; or an attempt at regulating the choice of words in familiar conversation*, appeared in 1794, in two volumes comprising over 900 (generously spaced) pages. In one sense, then, she could be placed among women using their learning as educators, producing glossaries and dictionaries for a domestic sphere. Piozzi herself suggests in her preface that her *Synonymy* should take its place on 'a parlour window, [...] unworthy of a place upon a library shelf' (Piozzi 1794, Vol. I, pp. iv–v). She appears to claim a space for women that does not impinge on male domains: 'while men teach to write with propriety, a woman may at a worst be qualified – through long practice – to direct the choice of phrases in familiar talk' (Piozzi 1794, Vol. I, p. iv). To give an example (Piozzi 1794: vol. II 9–11):

> **Malapert. Saucy, Impertinent.**
>
> THE last of these has by corruption become the common conversation word, and turned the first, which is the proper one, out of good company: for by IMPERTINENT is meant in strict propriety [...] the man goes to supper with his mistress when he hears she has an ague, and inveighs against the marriage stage when invited to celebrate a wedding dinner [...]. Now nothing of this perverseness is

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

> required to form what we are at present content to call IMPERTINENCE, falsely enough, for the MALAPERT miss, or SAUCY chambermaid, often possess sufficient skills to time their sprightly innocence and lively raillery reasonable well [...]. Whoever wishes to learn the full meaning of the word MALAPERT, may study the ready responses of an English miss, or an Italian chambermaid.

Russell argues that in focusing on word-choice in 'familiar conversation', Piozzi 'relocate[s] the lexicon from the abstract page to the concrete parlor, to (re)domesticate meaning in order to highlight its nuance, contingency and power, particularly in social circulation'. She perhaps thus even 'anticipates trends in lexicography that favour spoken corpora' (Russell 2018, pp. 147 f.).

Piozzi is, then, representative of much of what Russell has to say about women in lexicography. She is unusual in one way, however. While few women's contributions to English lexicography before the twentieth century were noteworthy for 'originating' something new, Russell judges that Piozzi's *British Synonymy* was ground-breaking, as the first of many publications in English inspired by Gabriel Girard's (1718) French synonymy, which were pre-cursors to Roget's famous *Thesaurus*, first published in 1852.

Russell also re-examines the roles played by women as paid (or more often unpaid) assistants, or as enablers of others' (men's) lexicographical activity by maintaining a household or providing companionship. The history of the iconic *OED* is rich in such stories. James Murray's daughters helped sort the slips on which attestations of words were written. Ada Murray, James Murray's wife, ran the household, reportedly had the idea to build an office (the so-called Scriptorium) for the dictionary in their home (Russell 2018, p. 150), and acted as Murray's unpaid secretary for many years. James Murray described Ada as 'the pivot on which the whole house revolved', and apparently consulted her on every important decision: it may have been at her urging that he took on the editorship of the *Dictionary* in the first place (Gilliver 2016).

Among the unpaid army of so-called readers around the country who recorded citations of words in use to be incorporated into the dictionary entries, there were by 1884 nearly 240 women. We can note, for example, Edith Thompson (1848–1929) and her sister Elizabeth (both authors in their own right). who contributed 15,000 quotations between 1880 and 1888, and continued through the rest of their lives too; Jennett Humphreys (1829–1917), a children's author, who had contributed nearly 20,000 quotations by 1888 (Gilliver 2016, n. p.).

Other women found a foothold as voluntary sub-editors. Five out of sixty sub-editors working on particular letters of the *OED* were women. Novelist Charlotte Yonge (1823–1901) was one of the first volunteer 'sub-editors', preparing draft entries in the letter N in the 1860s. One early paid member of staff was Ethelwyn Rebecca Steane (1873/4–1941), employed as an assistant by William Craigie, the OED's third editor, in 1901; she went on to work for the dictionary for three decades. Of course in the twentieth century, some woman forged a full lexicographical career within the *OED*. Jessie Senior (later Coulson) (1903–87) was among the first. She began work as an assistant in 1928, engaged in the compilation of the first Supplement to the *OED*, going on to establish a successful career as a lexicographer. The Shorter *OED*, the first edition of which appeared in February 1933, bore her name on the title page, the first such Oxford dictionary to do so, followed also a few months later by the Supplement to the *OED*. Amongst her work on other Oxford dictionaries, Coulson also compiled a Russian-English dictionary which appeared in 1975 (Gilliver 2016).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Reviewing Russell's book, Arias-Badia (2019) called for further work to examine the history of women in lexicography beyond English. As far as I know, similar systematic cataloguing of women's participation in lexicography remains a desideratum for other languages, and certainly for the German context with which I am most familiar. In this short paper, I cannot begin to approach Russell's meticulous documentation and incisive analysis of women's roles in the history of English lexicography, but I shall set out something of what we know already for some of the languages beyond English, before focusing on what we know for the field of German lexicography. Throughout, my approach is inspired by Russell – and my case studies chiefly lend weight to Russell's key conclusions. Women who found opportunities to undertake lexicographical work, even as late as the early twentieth century, were most likely to do so in domains that fell outside the interests of mainstream national-language monolingual lexicography: in domestic settings; in bilingual lexicography, including in missionary work; and in dialect lexicography. In some cases, though, their work was ground-breaking and of lasting significance.

2. Beyond English

The scale of the challenge set by Arias-Badia (2019) becomes clear when we turn to the important new edited volume on *Women in the History of Linguistics*, whose 19 chapters provide state-of-the-art overviews on women in the history of linguistics of European, African, American, Australian, and Asian languages. Even this fine volume yields very slim pickings for someone pursuing the history of women in lexicography, for it seems that Russell's bibliographic and analytical work in English lexicography has yet to be replicated for other languages.

A fifteenth-century Arabic source includes a tantalizing mention of a woman dictionary compiler, but no such dictionary has survived (Sadiqi 2020, p. 483). In Italy, Carolina Coronedi Berti (1869–1874) produced a two-volume dialect dictionary, *Vocabolario Bolognese Italiano* (Sanson 2020, pp. 84–86). In Spain, Luisa Lacal produced a *Diccionario de la musica tecnico, historico, bio-bibliografico* (1900 [1899]), while María Moliner prepared an unpublished dialect dictionary, and revised a Spanish dictionary published by the Spanish Royal Academy (1914) prior to a well-regarded dictionary of Spanish usage published in the 1960s (Calero Vaquera 2020, pp. 142 f.).

In Denmark, Margrethe Thiele (1868–1928), a practising scientific translator from Danish to French, pioneered work towards a Danish-French dictionary large enough to meet the needs of translators such as herself (Bull/Henrikson/Swan 2020, pp. 266 f.), going beyond the existing medium-size *Dansk-norsk-fransk Haand-Ordbog* of Sundby and Baruëls (1883–84). By 1910, Thiele had collected sufficient material to approach Jens Kristian Sandfeld (1873–1942) at Copenhagen University, himself involved in work on a dictionary of the Danish language (*Ordbog over det danske Sprog*; see Barr/Høybye 2014), and although the First World War delayed progress and access to funding, from 1918 onwards Thiele received an annual grant from the Carlsberg Foundation for her work. As illness slowed Thiele's progress, she involved Dr Andreas Blinkenberg from 1923, and after her death in 1928, he completed the dictionary and saw it through to publication in 1937 (Blinkenberg/Thiele 1937).

With its 1,700 double-column pages, the dictionary was the largest of its kind at the time, and the fact that Thiele was (like Blinkenberg) working out of her native language into a learned language makes it all the more impressive. Schøsler (2014) praises the systematic

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

structure of the dictionary's entries, which she suggests had an influence on later Danish lexicography: clear definitions and adequate translations in the first part of the entry, followed by exemplifications, collocations and expressions in the second part. To what extent these structural principles had been developed by Thiele is not clear. Thiele's material amounted to some individual 100,000 slips, which passed after her death to the Royal Library of Denmark, but that archive has not, it seems, received further attention.

The dictionary's entry for *kvinde* ('woman'), reproduced below, is noteworthy for how it attests to women's growing economic and social independence. We find phrases such as the 'fallen woman' and the Biblical injunction to be silent in church, but they are balanced by newer collocations that attest to the economic power of women, their independence in travel, their access to education, and their participation in competitions and in associations:

> Kvinde *c (-r)* femme*. ‖ *~rne (og.)* de la femme (*fx.* l'activité économique de la f., féminin (*fx.* le travail f.); *brav ~ f. de bien; falden ~ f. perdue; for ~r (paa Jærnbane) [Côté des] dames*; *Bogbindingsskole for ~r* école féminine de reliure; *Verdensmesterkab for ~r* championnnat féminin; *født av en ~* [un être né] de la femme; *løbe efter ~rne* aimer le cotillon; *Forening baade for Mænd og ~r* association mixte; *~n skal tie i Forsamlingen (bibl.)* que les femmes se taisent dans vos assemblées, que vos femmes se taisent dans les églises.

As Russell (2018, pp. 102–105) notes, one increasingly common form of bilingual lexicography by European women from the nineteenth century was undertaken as part of missionary or colonizing activity. To the examples identified by Russell up to the year 1900 (including, for example, Kilham 1820, a Wolof-English dictionary, and Woodward 1892, an English-Chichewa dictionary), we can add Daisy Bates (1859–1951), appointed by the Western Australian government to record word-lists for Kimberley languages in the early twentieth century (McGregor 2013), and Mary Haas (1910–1996), who produced dictionaries of two American Indian languages, Creek and Tunica (Heaton/Koller/Campbell 2020, pp. 356–358).

Another such missionary linguist, active in the early twentieth century, is Cinie Louw (1872–1935), who produced a two-way vocabulary of Karanga, a language spoken in southern Africa, as part of a language manual which also includes a grammar (Louw 1915). The English-to-Karanga part of the vocabulary (ibid., pp. 149–291) precedes and is almost half as long again as the Karanga-to-English part (ibid., pp. 291–395). This is, I suspect, somewhat unusual in the history of bilingual lexicography, where the target-language-to-source-language tends to be prioritized, and it possibly reflects the importance attached to ensuring that the authorized knowledge of the missionary/colonizer can be expressed in the local language. Louw (1915, pp. v-vi) explains that

> The Vocabulary of Part IV. does not claim to be either an exhaustive or correct dictionary. Such words have been collected as could be collected from the natives, and meanings assigned to them, which, it is hoped, will be found to be generally correct. […] I must also express my deep indebtedness to my faithful native helper Timotheus, who assisted me with untiring perseverance.

Despite the structural prioritization of English-to-Karanga, many of the entries under English headwords reveal how Louw's work is in fact shaped by how her informant supplies words. As the entries below, for *adulterer* and *woman*, show, Louw records relevant Karanga words even when there is no clear English lexeme for which they serve as equivalents. (Louw's numbers in brackets indicate the noun class).

> adulterer, adulteress, *mupati (1); mombḡe (4); nzenza (4); zeṉgeya (4).*
> very bad --- *mushwerakwenda (1); mvemveti (4); mbḡamati (4); ziṉveṉve (5).*

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

A 'very bad adulterer' is hardly a current English collocation or usual sub-entry, and its meaning is underspecified: we are left guessing as to what it distinguishes, as opposed to some more acceptable form of adultery. (In the other dictionary half, the word *mushwerak-wenda* is glossed as one who goes from one place to another.) However, the English phrase provides a 'slot' for recording Karanga words supplied by the informant. Similarly, some sub-entries under the lemma *woman* are in effect paraphrases for Karanga lexemes:

> woman, *mukadzi (1); munukadri (1).* --- who has borne children, *mvana (4)* --- whose children all die, *vumba (7), u ne pfuva (1)* lying-in ---, *muzere (1).* old --- *muchemgere (1)*; or *chembere (4).* a young married --- , *murovora (1).*---, a stranger who becomes the wife of a chief, *moromoka (4)*

Despite the English-to-Karanga format, then, once we progress beyond the initial headwords, the material of the entries often reads more like the result of a Karanga-to-English process, where English paraphrases are given for Karanga lexemes that Louw presumably felt deserved to be recorded, in order to capture the cultural specificities of the host society. To add further examples, under *aggravate*, we find 'aggravate illness by casting a shadow', and under *apportion*, 'apportion work in a garden'. Under the entry *assegai* (a light spear), we find Karanga terms for the spiral shaft, the wooden handle, the blade, edges, ridge, and point of such a spear. Similarly, headwords such as *apron, ant* and *antelope* have multiple equivalents in Karanga which are disambiguated through description in English (e. g. small black ant, large black ant; front and back aprons, aprons of men and of women).

3. Women in the history of German lexicography

So much, then, for the relatively few clues of women's early contributions to lexicography outside English that the histories in Ayres-Bennett/Sanson (2020) provide. I have no doubt that, just as for English, there are hundreds more, but the work to uncover them remains to be done. For the remainder of this paper, I shall focus on German, the language context with which I am most familiar. Loosely following Russell's approach, I shall consider women as imagined or actual users and dedicatees of dictionaries; women as unrecognized contributors to lexicographical works and as known compilers of dictionaries; and finally, very tentatively indeed, women as represented in dictionaries. My time-frame is, like Russell's limited to before about 1900, but including, like Ayres-Benett/Sanson (2020), women born before 1900 and active in the twentieth century.

3.1 Women invoked as imagined readers and as patrons

I noted above that Russell (2018, p. 41) identified 24 examples of English dictionaries before 1700 that named or invoked women, and that many of these works were multilingual. Knowledge of languages was an accomplishment 'intellectually appropriate for women and socially practical' (ibid., p. 47), and indeed could be essential for women of high social standing navigating international dynastic connections. Women were among the subscribers to John Minsheu's *Guide into Tongues* (1617), which includes German among one of several languages alongside English (Russell 2018, p. 38), but I am not aware of instances of German monolingual or multilingual dictionaries that invoke women in German before 1700. (There may well be some; that investigation has not been done). Nevertheless, we know that many German women with access to education were involved in language learning, and in translation, which – since it could be considered an exercise in language learning –

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

was one of the few acceptable ways for women to undertake scholarly work, even if the work usually remained unpublished. A handful of women, some of them practising poets, also became members of various language societies of the time (Brown 2009; McLelland 2020, pp. 196–200).

Women are explicitly invoked as an audience in the *Frauenzimmer–Gesprächsspiele* (*Conversation Games for Ladies*), published in eight volumes by Georg Harsdörffer between 1643 and 1649 and involving all kinds of language games. The contents are not lexicographic in any usual sense, but do contain some word lists, lists of emblems, and even a listing of hand sign language. The major German grammar published by Harsdörffer's contemporary and friend Justus Georg Schottelius (1612–1676) included lists of thousands of German root-words and their compounds (Schottelius 1663, pp. 1278–1446), intended as a basis for a future dictionary, much discussed within the language society of which he was a member, the Fruitbearing Society. Schottelius was also tutor to the children of his patron Augustus the Younger, Duke of Brunswick-Wolfenbüttel, including two daughters, for whom Schottelius wrote several plays, and he dedicated his poetics, first published in 1645, to their mother Elisabeth Sophie, herself a poet and musician. Might his interactions with them have influenced his lexicographical work in any way? We do not know.

The first known female dedicatee of a German dictionary is Sophia, Electoral Princess and duchess-dowager of Hanover, to whom Christian Ludwig dedicated his English-German dictionary, the very first bilingual dictionary of English and German, which he published in 1706, at a time of intensifying relations between the House of Hanover and England. Sophia, who conducted a substantial correspondence with Gottfried Leibniz, was known for her education and intelligence, 'long admir'd by all the Learned World, as a Woman of incomparable Knowledge in Divinity, Philosophy, History, and the Subjects of all sorts of Books, of which she has read a prodigious quantity' (Strickland 2011, p. 1, citing the philosopher and writer John Toland in 1705). Sophie was also heir to the thrones of England and Scotland (later Great Britain) and Ireland, though she died shortly before she would have become queen (so that her son succeeded her in 1714, as George I). She was, then, measured against what Russell has found for English, a prototypical female dedicatee, and especially for a multilingual dictionary: she was exceptionally highly educated and multilingual herself; and she was powerful. She was also interested in the instruction of her children, preparing for life as English royalty, so that Ludwig's dictionary was likely to be of practical value too. In the following century, the 1846 edition of Hilpert's bilingual German-English dictionary is likewise dedicated to an important Hanoverian woman, Queen Victoria (as the *OED* would later be), jointly with her German-born and German-educated husband Albert, whom she had married in 1840. Queen Victoria's mother was German, she had had a German governess, and she and Albert employed a German governess for their children.

3.2 Women as contributors to lexicographical work

As for the hidden role that women may have played in dictionary-making, we can do no more than speculate on whether and how household members of known male dictionary compilers might have supported that work. Caspar Stieler, who compiled the first complete dictionary of German, *Der teutschen Sprache Stammbaum und Fortwachs* (1691), was married twice – did either of his wives Regina and Christiane Margarethe Cotta have any involvement in the dictionary, or was their role restricted to running the household that enabled Stieler to complete his task? Again, we do not know.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Women who participated directly in German lexicographical work are hard to identify before the late nineteenth century. There are no women at all amongst the seventy-eight authors listed in Moulin-Fankhänel's two-part bibliography (1994, 1997) of German grammars and orthographies up to the end of the seventeenth century, nor in William J. Jones's (2000) bibliography of seventeenth-century German lexicography. Even in the area of language purism – one of the most widespread forms of lay engagement with linguistic ideas, and a prominent thread in the history of the German language from about 1500 – there is only one woman represented amongst the 117 texts in Jones's (1995) documentation of foreign word purism between 1478 and 1750, and it is not a lexicographical text. It may be that there are instances of dictionary-like material in so-called *Anstandsliteratur* (manners guides) and letter-writing guides written by and/or for women. Certainly women's language was a topic in some of these works, including in works written by women (see McLelland 2020, pp. 200–203).

Luise Gottsched, née Kulmus (1713–1762), wife of Johann Christoph Gottsched (1700–1766), seems to have enjoyed working with her husband, rather than caught in 'literary drudgery work' for him (Lerner, cited in Brown 2012, p. 3). However, the Gottscheds' activities did not include lexicographical work. As for the Grimms and their *Deutsches Wörterbuch* project, we might expect that, as with the *OED*, women played a role 'behind the scenes'. Lelke (2005) examined women's contribution to the work of the Grimm brothers, devoting a chapter to 'women in the house' (Lelke 2005, pp. 190–250). Her study shows how, through participation in that half-private, half-public intellectual world, women like Wilhelm Grimm's wife Dorothea Grimm and others including Bettina von Arnim and Sarah Austin had some scope to contribute to the academic work of the Grimms and their circle, but it seems their work did not progress beyond assisting in or helping publicize the work of men.

3.3 Women lexicographers in German

3.3.1 Elizabeth Weir

It was not a German, but an English woman who is, as far as I can see, the first named woman who produced a German dictionary: Elizabeth Weir. That is perhaps no surprise, given the pattern identified by Russell – and largely borne out by my few examples beyond English – that the mainstream work of monolingual lexicography remained out of reach of women before the twentieth century. Weir's bilingual German-English dictionary appeared in 1888 as *Heath's New German Dictionary* in Boston (and as *Cassell's New German Dictionary* in Britain). Frustratingly, despite careful detective work by Husbands (2001), nothing is known of Elizabeth Weir beyond what her preface reveals, written while she was living in Stuttgart in 1888, where the second, English-German part was largely written, and where, she reports, German friends helped her with numerous technical expressions and idioms that, 'though of common occurrence in every-day life, are not generally found in dictionaries'.

Our lack of knowledge about Weir's background and training is particularly frustrating because Weir's original contribution seems to have been substantial. Weir's dictionary was intended to serve the 'young student' as 'a handy volume', with 'a collection of idioms, proverbs, and quotations [...], which is larger and more varied' than in other dictionaries' (Weir 1888, p. v). Virtually all of the preceding dictionaries had been compiled by Germans and intended for German learners of English. This meant, Weir explained, that they had 'not provided for the difficulty which the English student feels when called to select from some

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

dozen German words the special one which answers to the special sense in which the English word is to be used' (Weir 1888, p. v). Weir's aim, by contrast, was to produce a dictionary really suitable for English learners of German, and her dictionary is the first to give disambiguations of different senses in English rather than German. Among her predecessors, even the 1841 revised edition of Flügel's dictionary, ostensibly 'adapted to the English student' (as the title page states), had not yet done this. The first few lines of Weir's entry for *Head* show what this looks like in practice, as Weir's paraphrases in English allow the English student to select the appropriate German equivalent:

> Head, I. s. das Haupt, der Kopf; *(individual)* das Individuum, der Mann, das Stück; *(chief)* das Haupt, der Häuptling, Führer; *(principal)* der Vorsteher, Verwalter, Direktor; *(chief place)* das Haupt, die Spitze; *(understanding)* der Kopf, Verstand; *(prow)* der Schiffschnabel; *(source)* die Quelle; die Höhe, Krisis *(of an illness)*; *(division)* der Punkt, Hauptpunkt, Abschnitt, Paragraph [...][1]

Weir's dictionary is more concise than that of her predecessors – both dictionary halves fit into a single 'handy volume' – but Weir still made a particular effort to give plentiful examples of how words are used in context 'thoroughly illustrative of the points in the two languages in which they differ from one another' (Weir 1888, p. v). For instance, under *head*, we find examples where a literal translation of 'head' will not do:

> To make neither ---- nor tail of, aus (einer Sache) nicht klug werden können; [...] ---- of the stairs, der oberste Theil einer Treppe; [...] she sat at the ---- of the table, sie saß oben am Tische

The last example – where a woman sits at the head of the table – likewise stands out in contrast to examples given by Weir's predecessors under the same lemma, where none of the people taking a position as head or at the head of something is a woman. By contrast, indeed, Hilpert (1857) gives *The husband is the ---- of the wife, der Mann ist des Weibes Haupt.* Whether Weir's introduction of female headship is a single isolated example or perhaps representative of a more systematic approach by Weir remains to be investigated. Taken together with the example of Thiele's treatment of the headword *kvinde*, discussed above, it hints tantalizingly that women lexicographers produced different dictionaries to men. More detailed study of the dictionaries of such early women lexicographers also has the potential to add a historical dimension to more recent debates about the extent to which dictionaries may perpetuate gender stereotypes, something the pioneering feminist linguist Pusch (1984) showed in her witty analysis of the *DUDEN-Bedeutungswörterbuch* (1970) as a story with disappointingly marginal and feeble female characters.

Weir's work was clearly considered successful, for the prominent Germanist Karl Breul, the first Schroeder Professor of German at Cambridge, undertook to produce a revised version of it. When it appeared in 1906, Breul thanked his former students 'the Misses G. M. Parry, H. Sollas, and J. Burne' (Breul/Weir 1906, pp. v–vi), and above all 'Miss Minna Steele Smith, Head Lecturer in Modern Languages at Newnham College, Cambridge', who assisted in checking the proofs. These women's roles conform to the pattern that Russell identified of women as assistants rather than protagonists in the business of dictionary-making in the nineteenth century. This makes the gap in our knowledge about Elizabeth Weir, whose work underpins Schroeder's later edition, all the more frustrating.

[1] Here, and in examples from other dictionaries below, I have not attempted to replicate the use of different fonts (black letter and antiqua), used for German and English respectively.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

3.3.2 Klara Hechtenberg Collitz

Klara Hechtenberg Collitz (c. 1865–1944) is another women lexicographer of German who was a partial outsider at least. Born in Germany, she trained as a teacher, then studied French in Lausanne, then English at the University of London and Oxford, and taught in Belfast and in America, before returning to study in Germany, gaining her PhD from the University of Heidelberg in 1901, returning to Oxford University as a lecturer in German (1901–1904). She then married and moved to America, and did not hold an academic position again, but she continued to research, and her publications include an alphabetical *Fremdwörterbuch des 17. Jahrhunderts* ('Foreign-word dictionary of the seventeenth century', 1904), with a list of 3380 foreign words, and *Verbs of Motion in their Semantic Divergence* (1931), which contained alphabetical listings of verbs of motion in Greek, Latin, German, English, French, Italian, and Spanish with analysis of their figurative use with senses of 'propriety, fitness, suitability, or related meanings' (Collitz 1931, p. 7; see Maas 2018; McLelland 2020, pp. 211 f.).

A generation after Hechtenberg Collitz, the first two women trained entirely within the German-speaking world who had careers as lexicographers are both already well known today for their work: Agathe Lasch and Luise Berthold.

3.3.3 Agathe Lasch

Agathe Lasch (1879–1942) was the first woman to follow a conventional academic path in German linguistics. After gaining a PhD from Heidelberg and then her habilitation from Hamburg in 1919, where she was initially a postdoctoral assistant to Professor Conrad Borchling, Lasch was in 1923 given a so-called extraordinary chair in Low German philology, thus becoming the first German *Professorin* in Germany (though the 'extraordinary' title in effect meant the rank of professor without the funds for assistants and support that go with a chair in the German system). Lasch had already published an important grammar of Low German in 1914; in 1917, while still a postdoctoral assistant to Borchling, Lasch was given the role of running a newly established dictionary archive. In this role, she was responsible for planning and collecting material for a dictionary of the variety of German spoken in the city of Hamburg, *Hamburgisch*. The dictionary was ground-breaking, not just in recording a city vernacular rather than a rural dialect, but also because Lasch used both systematic evaluation of historical sources, and questionnaires to capture current Low German usage, yielding 180,000 attestations by 1933. Lasch was in effect taking a sociolinguistic approach to dialectology to capture the changing status, and heterogeneity of, Low German in Hamburg, past and present (Schroeder 2009, p. 49). The dictionary of *Hamburgisch* was completed in 2006, still following the basic structure devised by Lasch (ibid., p. 47).

In 1923, Lasch, now a professor herself, launched a second major lexicographical project, a concise dictionary (*Handwörterbuch*) of Middle Low German, finally completed in 2009 (Schroeder 2009, pp. 56–58). Lasch again devised the structural framework to be followed, and also worked on seven fascicles of the dictionary herself. A concise dictionary could not include examples of words in context, or information on the temporal and regional distribution of individual words, as Lasch would have liked if space had allowed. Nevertheless, it benefited from recent work on the Middle Low German vowel system that had in part been triggered by Lasch's Low German grammar. For example, Umlaut was systematically marked, and original long vowels were distinguished from long vowels that were the result of vowel lengthening (Schroeder 2009, p. 58).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Lasch, a Jew, was forced out of her post and into 'retirement' in 1934. After unsuccessful attempts to emigrate, she was deported in 1942, and was killed in Riga in the same year (Kaiser 2009, p. 21). Despite her tragically curtailed career, she had a decisive impact on Low German lexicography.

3.3.4 Luise Berthold

Luise Berthold (1891–1983) is second only to Agathe Lasch in her pioneering role as a woman in German lexicography, again in German dialectology. Berthold studied German philology at Marburg and then, alongside her doctorate (awarded 1918), devoted half her time to working on the Hessisch-Nassau dialect dictionary, funded by the Prussian Academy of Sciences. The first fascicle of the *Hessen-Nassauisches Volkswörterbuch* was published in 1927, and in 1930 Berthold was, like Lasch, made an extraordinary professor, though she was awarded a full chair only in 1952 (Berthold 2008, pp. 110 f.).

The Hessen-Nassau dictionary, the compilation of which Berthold led from 1934, stood in the tradition of the Marburg school of dialectology, specifically Wenker's *Sprachatlas*. Just as Georg Wenker had used questionnaires to gather data to map the geography of sound changes in the nineteenth century, so Berthold proposed a new series of questionnaires that would yield word-geography maps for the dictionary (Berthold 2008, p. 53), an approach which became a model for later work. Both the Prussian and Mecklenburg dialect dictionary projects, which both began publication in 1934, followed the example of using word-geography maps, as did the German Word Atlas project itself (*Deutscher Wortatlas*, ed. W. Mitzka et al., 1951–1980), which Berthold was in charge of for a time after World War II.

3.4 Women in German/English bilingual dictionaries

Russell (2018) devotes her final chapter to feminist lexicography, one dimension of which has been the uncovering of the systemic ways in which definitions and examples have under-represented, stereotyped, or misrepresented women.[2] Of course, given what we know of the history of power relations, what we are likely to find is predictable. Russell (2018, p. 184), citing the provocative title of a short piece, 'Women are alcoholics and drug addicts, says dictionary' (Kaye 1988), noted drily that by 1988, such a finding should hardly have been surprising. Russell also warns that analysing '"isolated instances of ideological bias in definitional text" does very little to enrich our understanding of the inevitable partiality of lexicography' (Russell 2018, p. 174, citing Ogilvie 2013, p. 86). Nevertheless, there is still a case to be made for providing evidence and for bearing witness to the phenomenon, and arguably doing so is all the more valuable when examining historical sources, thus complementing the social history and discursive histories of gender, sexuality, and minoritization. The representation of women, of sexuality, and of minoritized groups, has accordingly come under scrutiny in recent work on history of English lexicography (e. g. Iamartino 2010, Turton 2019; see also Brewer 2005–). I shall end this paper, then, with a very preliminary exploration of two words in the field of sexuality and gender in a group of dictionaries that I have been looking at for a different project: bilingual German/English dictionaries of the eighteenth and nineteenth centuries. These dictionaries have received very little attention to

[2] One might also explore the representation of women among the authors from whom citations are taken. To what extent past German lexicography has represented or under-represented woman authors in its attestations is, as far as I know, also uncharted territory.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

date.[3] To explore this theme adequately would therefore be a major undertaking, and here I present case studies of just two words, with no claim to generalizability, but given as indication of how such a project might be rewarding: the German headword *Hure* ('whore'), and the English headword 'woman'.

3.4.1 German *Hure*

My first case study is the word *Hure*, 'whore', a word where the German and English words are cognate and have broadly similar scope. I was initially curious to see how Ludwig (1716), the first producer of a German to English dictionary, rendered *Hure* in English, the first time it ever needed to be done in a dictionary. It was somewhat unexpected to find that Ludwig gives fully ten equivalents in English, with no immediate further differentiation:

> **Hur oder Hure (die)** a whore, wench, harlot, prostitute, strumpet, crack, cucquean, trull, cockatrice, doxy. [...]

The explanation for this richness is disappointingly prosaic, however. In 1706 Ludwig had published an English–German dictionary, based on two earlier bidirectional French-English dictionaries by Abel Boyer (1699, 1700). In Boyer (1699), Ludwig would have found the following entry:

> **PUTAIN, S.F.** (*Fille ou femme prostituée*) Whore, Wench, Harlot, Prostitute, Strumpet, Crack, Cockatrice, Doxy

Eight of Ludwig's ten equivalents come, then, straight from Boyer's equivalents for *putain,* in the same order. The remaining two, *cucquean* and *trull*, are both listed as English headwords by Boyer, and following him, by Ludwig (1706). In each case Boyer gives *Putain* as one of the possible equivalents, and *eine hure* is the only German equivalent that Ludwig gives. (Ludwig indeed gives *hure* as an equivalent for all ten English terms, but often among others.) There is no mystery, then, in how Ludwig arrived at the English equivalents for *Hure* in his pioneering dictionary, and there is nothing to say about how he differentiates them. He does not.[4]

What about Ludwig's successors in the German-English lexicographical tradition? The first competitor to Ludwig, Theodor Arnold (1753), lists the same ten items as Ludwig, and in the same order, except that *cockatrice* and *doxy* are reversed:

> *Hure*, a Whore, Wench, Harlot, Prostitute, Strumpet, Crack, Wench, Cucquean, Trull, Doxy, Cockatrice

At the very end of the eighteenth century, a later edition of Arnold's dictionary (Bailey/Fahrenkrüger/Arnold 1797) and Ebers (1796–99) both still offer the same list of ten terms. There is, then, virtually no change over almost a century in the equivalents given, though the 1797 dictionary adds *drab*, and, more significantly, Ebers (1796–99) also adds three euphemistic terms *A woman of the Town, a Woman of Pleasure, a Courtezan.*

---

[3] Stein's (1985) survey stops with Ludwig (1706). Hartmann (2007) includes Ludwig (1706) and Flügel (1838), and Adler (1848), the latter in fact closely based on the revised edition of Flügel (1841). Cormier (2009) briefly discusses Ludwig, and mentions Theodor Arnold, Johann Christoph Adelung, and Johannes Ebers, on whom see also Lewis (2013).

[4] Note also the equivalents given for compounds with *-hur* later in the same entry:'**Eine schand-hur, soldaten-hur, allermans-hur, allgemeine hure** a prostitute, tomboy, drab, camp-whore, romp, rig, slut, jade or wench; a common whore, a common hackney'. The term *tomboy* here is presumably intended in the now obsolete sense of 'forward, immodest, or unchaste woman' (*OED* online, s. v. *tomboy*).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

In the nineteenth century, the compilers of a revised edition of Flügel (1841) stated their intention to refresh the dictionary while also attending to propriety. Flügel's original dictionary was, they judged, full of unnecessary and unsuitable material, in which 'the forgotten obscenities of the 17th and 18th centuries have been raked together into one heap' (Flügel 1841, p. iii). The revised 1841 dictionary accordingly lists just four English equivalents for *Hure*: *whore, harlot, strumpet, prostitute.*

Some years later, the professed aim of Hilpert (1857) was to give

> the most modern and the most colloquial forms to its expressions, instead (as been heretofore almost universally the case with such German and English dictionaries) of copying and handing down from lexicon to lexicon old terms and forms of speech' (Hilpert 1857, p. xv).

Hilpert (1857) gives the same four English equivalents for *Hure* (*whore, harlot, strumpet, prostitute*), together with two euphemisms in English (*a common woman, a woman of the town*, the latter already found in Ebers 1796–99). Lucas (1868) keeps largely the same terms as Hilpert (1857), *whore, harlot, prostitute, strumpet, woman of the town*, but also introduces another euphemism, *street-walker* in 'zur --- werden, to turn prostitute, to turn street-walker', the first time *street-walker* is included for *Hure*, although it was already included as an English headword by Ludwig (1706), glossed there as *eine gassenhure.*

There is, then, little evidence of a sensitivity to connotations of the different English terms for women who sell sex in Ludwig (1716) and his successors into the mid-nineteenth century. However, the inclusion, from the late eighteenth century onwards, of euphemistic English equivalents for *Hure* is noteworthy, and perhaps needs to be considered as part of an emerging wider sensitivity to vulgarity and obscenity – something we have seen was indeed explicitly thematized by the revisers of Flügel (1841). A related development is that in Flügel (1841), we are warned about the 21 noun *huren-* compounds listed: 'these are with a few exceptions, all vulgar', the first such warning in this lexicographical tradition for *Hure* (even though Ludwig did use such metalinguistic labelling when he chose to). In Hilpert (1857), the German base term *Hure* is itself now marked † for 'vulgar'.[5]

We can also detect a subtle change in how the German term is understood. Bailey/Fahrenkrüger/Arnold (1797) was the first to differentiate two figurative usages for *Hure* (marked *f.* below) to indicate that the term may be used, in an extended sense, of any woman caught in unchaste behaviour:

> [...] *f. eine geschwächte Person* defloured [sic] maid, lady; *f. jede weibliche Person, welche die Keuschheit oder eheliche Treue verletzt* lady -- woman of pleasure, one of the family of love

The new distinction of a separate figurative sense for *Hure* made by Bailey/Fahrenkrüger/Arnold (1797) is almost certainly taken from Adelung's monolingual German dictionary (1793), which distinguishes first the narrow sense, then two wider senses, which apply either to an unmarried woman who has become pregnant (a use 'in der harten Sprechart und im gemeinen Leben') or to any woman, whether married or not, 'welche durch unerlaubten Beyschlaf die Keuschheit verletzet, gleichfalls nur in der harten Schreibart und mit beleidigender Verachtung'.

[5] Probably following Heyse (1833) in the monolingual German tradition.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Hilpert (1857) followed suit, but now gave the figurative sense for *Hure* first, as had the more recent German dictionary of Christian Heyse (1833).

> 1) [in general] any woman who violates chastity [*eine gefallene*]. *Ein Mädchen zur --- machen*, to debauch or deflour [sic] a girl; *zur --- werden*, to become or be defloured or debauched; *sie hat ihre Tochter selbst zur --- gemacht*, she has prostituted her daughter herself.
>
> 2) [in a more limited sense] a woman who prostitutes her body for hire, a harlot, prostitute, a whore, a common woman, a woman of the town, a strumpet.

Over a period of some hundred and fifty years of German-English lexicography, then, even though the English equivalents change little, we see a changing sensitivity to the acceptability of the term *Hure*; an emergence of euphemistic language; and a sensitivity to the idea that there is a distinction to be preserved between a woman who actually sells sex for money and one who is willing to have sex with a man outside of marriage.

3.4.2 English *woman*

My second exploration concerns entries under the English headword *woman*. Ludwig (1706) gives a very simple entry:

> *Woman*, eine frau, ein weib, femme, *A lady's woman, a waiting woman*, einer damen kammerfrau, la femme de chambre d'une dame. *A woman of the town*, ein unzüchtiges weib, eine hure, *une femme debauchée, une putain.*

Arnold (1752) gives a far fuller entry than Ludwig for *woman*, with several idioms, which are, as far as I can tell, his own selection:

> WO'MAN, (wumän, V. S. wiman, prob. V. wamb u. Man) *femme*, mulier, fœmina, das Weib, die Frau. WOMEN, Money and Wine, have their good, and their Ruin, *femmes, argent & vin, ont leur bien et leur venin*, in muliere, pecunia, et vino venenum, Weiber, Geld und Wein, pflegen so schädlich als nutzlich zu seyn. Three WOMEN and a Goose make a Market, *deux femmes font un plaid, trois un grand Caquet, quatre un plein marché,* est quasi grande forum, vox alta trium mulierum, drey Weiber und eine Gans machen einen Jahrmarkt. The more WOMEN look in their Glasses, the less they look to their Houses, *femme qui trop se mire, peu file*, quæ in speculo diutius seipsam intuetur, colum neglegit et fusum, je fleißiger die Weiber in Spiegel sehen, je weniger sehen sie nach ihrer Haushaltung. WOMEN laugh when they can, and weep when they will, *femme rit, quand elle peut, et pleure, quand elle veut*, quoties potest ridet, stet autem quando lubet mulier, die Weiber lachen, wenn sie können, und weinen, wenn sie wollen. A WOMAN conceals what she knows not, *une femme cache ce qu'elle ignore*, quod nescit fœmina, celat, eine Frau verschweigt, was sie nicht weiß. Tell a WOMAN she's handsome, but once, the Devil will tell her so fifty times, *dis à une femme, qu'elle est belle, et le diable lui le dira cinquante fois*, pulchritudo nimis laudata tumescit, wenn man das Frauenzimmer gar zu sehr lobet, wird es nur stolz.

Arnold's entry is an eloquent instance of all that feminists have objected to in dictionary-making by men. From the six idioms that Arnold gives, it emerges that i) women – likened to consumables, money and wine – can lead to ruin; ii) women are overly talkative and loud, so that three together is like a market; iii) women are vain, and likely to neglect their domestic duties; and iv) women are deceptive, able to weep on demand, and adept at concealing their ignorance.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

The impact of these depictions of womanhood on the reader is arguably all the more emphatic for being repeated in four languages, English, German, French and Latin. Further work would be needed to determine if this misogynist selection of material is typical of Arnold, or simply an isolated instance in his work. It does not seem to have set the direction for future English–German bilingual lexicography, at any rate. Adelung (1783/1793) based his work on Johnson (1755)'s English dictionary, which gives literary attestations of use. From Johnson's nine attestations for the word *woman* – from Shakespeare, the Bible, and other sources – Adelung selects just two, one admittedly stereotyping from Addison ('Vivacity is the gift of women, gravity that of men') and one illustrating the use of *woman* to refer to a female servant to a lady ('By her woman I sent your message', Shakespeare).

4. Conclusion

This paper began with an overview of what is currently known about women in the history of lexicography. With the exception of the exemplary work of Russell (2018) for the case of English, this remains largely uncharted territory, and for languages other than German, I have done little more here than look a little more closely at the two instances identified mentioned in Ayres-Bennett/Sanson (2020) that were accessible to me: Thiele's work towards a comprehensive Danish–French dictionary and Louw's (1915) vocabulary of Karanga.

As for the history of women in German lexicography, again much more needs to be done, but what we know thus far suggests a similar pattern to that identified by Russell of women as patrons and dedicatees, but also of participation by women outside the mainstream of national dictionary-making, at least as far as the early twentieth century: in particular in the spaces afforded them in bilingual lexicography (Weir), in lexicographical projects that supplement mainstream dictionaries (Collitz's foreign-word dictionary) and in the area of dialectology (Lasch, Berthold). It is worth emphasizing the importance of these works, however: Weir's dictionary was successful and innovative; Collitz's foreign-word dictionary is still included on reading lists today; and, in the twentieth century, both Lasch and Berthold took charge of important lexicographical projects that were pioneering in method and far-reaching in their influence.

References

Adelung, J. C. (1783): Neues grammatisch-kritisches Wörterbuch der englischen Sprache für die Deutschen vornemlich aus dem größern englischen Werke des Hr. Samuel Johnson […]. Leipzig.

Adelung, J. C. (1793): Grammatisches-kritisches Wörterbuch der Hochdeutsche Mundart [...]. Leipzig.

Adler, G. (1848): Dictionary of German and English languages. New York.

Arias-Badia, B. (2019): [Review] Lindsay Rose Russell. 2018. Women and dictionary making: gender, genre, and English language lexicography. In: International Journal of Lexicography 32 (3), pp. 389–391.

Arnold, T. (1753): Neues Deutsches-Englisches Wörterbuch [...]. Neue verbesserte Auflage. Leipzig.

Arnold, T./Bailey, N. d. (1752): A compleat English dictionary: oder vollständiges englisch-deutsches Wörterbuch. Leipzig.

Ayres-Bennett, W./Sanson, H. (2020): Women in the history of linguistics. Oxford.

Bailey, N./ Fahrenkrüger, J. A./Arnold, T. (1797): Nathan Bailey dictionary English-Geman and German-English oder Englisch-Deutsches und Deutsch-Englisches Wörterbuch. Leizpig.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/ Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Barr, K./Høybye, P. (2014): Kr. Sandfeld (Jens Kristian Sandfeld). In: Dansk Biografisk Leksikon. https://biografiskleksikon.lex.dk/Kr._Sandfeld (last access: 12-04-2022).

Berthold, L. (1927–): Hessen-Nassauisches Volkswörterbuch. Marburg.

Berthold, L. (2008): Erlebtes und Erkämpftes. Rückblick einer Pionierin der Alma Mater. Königstein/Taunus. [First self-published by the author, 1969].

Blinkenberg, A./Høybye, P. (1964–1966): Fransk-dansk ordbog I-II. Copenhagen.

Blinkenberg, A./Thiele, M. (1937): Dansk-fransk ordbog. Copenhagen.

Boyer, A. (1699): The royal dictionary. In two parts, first French and English. Secondly, English and French. […]. London.

Boyer, A. (1700): The royal dictionary abridged. In two parts. I. French and English. II. English and French [...]. London.

Breul, K./Weir, E. (1906): A new German and English dictionary. Revised by K. Breul. [S. I.].

Brewer, C. (2005–): Examining the OED. https://oed.hertford.ox.ac.uk/ (last access: 12-04-2022).

Brown, H. (2009): Women translators in the Sprachgesellschaften. In: Daphnis 38, pp. 621–647.

Brown, H. (2012): Luise Gottsched the translator. Rochester/New York.

Bull, T./Henriksen, C./Swan, T. (2020): Obstacles and opportunities for women linguists in Scandinavia. In: Ayres-Bennett, W./Sanson, H. (eds.): Women in the history of linguistics. Oxford, pp. 245–278.

Calero Vaquera, M. L. (2020): The contribution of women to the Spanish linguistic tradition: Four centuries of surviving words. In: Ayres-Bennett, W./Sanson, H. (eds.): Women in the history of linguistics. Oxford, pp. 91–120.

Collitz, K. H. (1931): Verbs of motion in their semantic divergence. Philadelphia.

Cormier, M. C. (2009): Bilingual dictionaries of the late seventeenth and eighteenth centuries. In: Considine, J. (ed.): The Oxford history of English lexicography. Oxford, pp. 65–85.

Ebers, J. (1796–99): The new and complete dictionary of the German and English languages composed chiefly after the German dictionaries of Mr. Adelung and of Mr. Schwan. Leipzig.

Evelyn, M. (1690): Mundus muliebris: or, the ladies dressing-room unlocked, together with the fop-Dictionary, Compiled for the use of the fair sex. London.

Flügel, J. G. (1830, 1838, 1841): A complete dictionary of the English and German and German and English languages. Leipsic. Second ed. 'improved & augmented', appeared Leipsic, 1838; 1841 ed. 'adapted to the English student, with great additions and improvements, by C. A. Fieling and A. Heimann'.

Girard, G. (1718): La Justesse de langue Françoise ou les Différentes Significations des mots qui passent pour synonymes [...]. Paris.

Harsdörffer, G. P. (1968–1969 [1643–1649]): Frauenzimmer Gesprächsspiele. Endter. Reprint, ed. Irmgard Böttcher. Tübingen.

Hartmann, R. R. K. (2007): Interlingual lexicography: selected essays on translation equivalence, contrastive linguistics and the bilingual dictionary. Munich.

Heaton, R/Koller, E./Campbell, L. (2020): Women's contribution to early American Indian linguistics. In: Ayres-Bennett, W./Sanson, H. (eds.): Women in the history of linguistics. Oxford, pp. 345–366.

Hechtenberg, C. (1904): Fremdwörterbuch des 17. Jahrhunderts ('Foreign-word dictionary of the seventeenth century). Berlin.

Heyse, C. (1833): Handwörterbuch der deutschen sprache […]. Magdeburg.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Hilpert, J. L. (1828–33, 1857): A dictionary of the English and German languages. Carlsruhe. [New cheaper edition Leipzig, 1857.]

Husbands, C. T. (2001): Who was Elizabeth P. Weir?: Gender visibility and female invisibility in the world of lexicography. In: The Linguist 40 (2), pp. 48–51.

Iamartino, G. (2010): Words by women, words on women in Samuel Johnson's dictionary of the English language. In: Considine, J. (ed): Adventuring in dictionaries: new studies in the history of lexicography. Cambridge, pp. 94–124.

Johnson, S. (1755): A dictionary of the English language […]. In two volumes. London.

Jones, W. J. (1995): Sprachhelden und Sprachverderber (1478–1750). Berlin.

Jones, W. J. (2000): German lexicography in the European context: a descriptive bibliography of printed dictionaries and word lists containing German language (1600–1700). Berlin.

Kaiser, C. M. (2009): Zwischen »Hoffen« und »Verzagen« Die Emigrationsbemühungen Agathe Laschs. Ein Werkstattbericht. In: Nottscheid, M./Kaiser, C. M./Stuhlmann, A. (eds.): Die Germanistin AGATHE LASCH (1879–1942). Aufsätze zu Leben, Werk und Wirkung. Nordhausen, pp. 11–46.

Kaye, P. (1988): Women arre alcoholics and drug addicts, says dictionary. In: English Language Teaching 43 (3), pp. 192–195.

Kilham, H. S. (1820): Ta-re wa-loof. Ta-re boo Juk-à. First lessons in Jaloof. Tottenham.

Lasch, A./Borchling, C. (1956): Mittelniederdeutsches Handwörterbuch. Completed by Gerhard Cordes (1956). Neumünster.

Lelke, I. (2005): Die Brüder Grimm in Berlin: zum Verhältnis von Geselligkeit, Arbeitsweise und Disziplingenese im 19. Jahrhundert. Frankfurt.

Lewis, D. (2013): Die Wörterbücher von Johannes Ebers. Studien zur frühen englisch-deutschen Lexikographie. Dissertation. Würzburg.

Louw, C. (1915): A Manual of he Chikaranga Language, with grammar, exercises and useful conversation sentences and vocabulary: English Chikaranga and Chikaranga-English. Bulawayo.

Lucas, N. I. (1854–68): Englisch-deutsches und deutsch-englisches Wörterbuch: mit besonderer Rücksicht auf den gegenwärtigen Standpunkt der Literatur und Wissenschaft. Bremen.

Ludwig, C. (1706): A dictionary English, German, and French […]. Leipzig.

Ludwig, C. (1716): Teutsch-Englisches Lexicon […]. Leipzig.

McGregor, W. B. (2013): Daisy Bates' documentations of kimberley languages. In: Language & History 54 (2), pp. 79–101.

McLelland, N. (2020): Women in the history of German language studies: 'That subtle influence for which women are best suited'. In: Ayres-Bennett, W./Sanson, H. (eds.): Women in the history of linguistics. Oxford, pp. 193–217.

Minsheu, J. (1617, rpt. 1978): Ductor in linguas = Guide into the tongues […]. Delmar.

Mitzka, W. et al. (1951–1980): Deutscher Wortatlas. 22 vols. Giessen.

Moulin-Fankhänel, C. (1994, 1997): Bibliographie der deutschen Grammatiken und Orthographielehren. I. Von den Anfängen der Überlieferung bis zum Ende des 16. Jahrhunderts. Bd. II. Das 17. Jahrhundert. Heidelberg.

Murray, J. (1900): The evolution of English lexicography. The romanes lecture delivered in the Sheldonian Theatre, Oxford, June 22, 1900. Oxford.

Ogilvie, S. (2013): Words of the world: a global history of the Oxford English Dictionary. Oxford.

Pasch, H. (2020): European women and the description and teaching of African languages. In: Ayres-Bennett, W./Sanson, H. (eds.): Women in the history of linguistics. Oxford, pp. 487–508.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Piozzi, H. L. (1786): Anecodotes of the late Samuel Johnson, L. L. D. during the last twenty years of his life. Dublin.

Piozzi, H. L. (1794): British synonymy: or, an attempt at regulating the choice of words in familiar conversation. Dublin.

Pusch, L. (1984): »Sie sah zu ihm auf wie zu einem Gott«: das DUDEN-Bedeutungswörterbuch als Trivialroman. In: Weiblichkeit oder Feminismus? Beiträge zur interdisziplinären Frauentagung, Konstanz 1983. Konstanz, pp. 57–66.

Reddick, A. H. (1990): The making of Johnson's Dicionary, 1746–1773. Cambridge.

Roget, P. M. (1852): Thesaurus of English words and phrases [...]. London.

Russell, L. R. (2018): Women and dictionary making: gender, genre, and English language lexicography. Cambridge.

Sadiqi, F. (2020): Women and the codification and stabilization of the Arabic language. In: Ayres-Bennett, W./Sanson, H. (eds.): Women in the history of linguistics. Oxford, pp. 469–486.

Sanson, H. (2020): Women and language codification in Italy: marginalized voices, forgotten contributions. In: Ayres-Bennett, W./Sanson, H. (eds.): Women in the history of linguistics. Oxford, pp. 59–90.

Schøsler, L. (2014): Blinkenberg og Høybyes Dansk-Fransk og Fransk-Dansk Ordbog – fra seddelsamling til trykte ordbøger og til onlineudgave. In: LexicoNordica 21, pp. 161–180.

Schottelius, J. G. (1663): Ausführliche Arbeit von der teutschen Haubtsprache. Braunschweig: Zilliger. Rpt. ed. Wolfgang Hecht. Tübingen, 1967.

Schroeder, I. (2009): Agathe Lasch und die Hamburger Lexikographie. In: Nottscheid, M./Kaiser, C. M./Stuhlmann, A. (eds.): Die Germanistin AGATHE LASCH (1879–1942). Aufsätze zu Leben, Werk und Wirkung. Nordhausen, pp. 47–62.

Stein, G. (1985): Englisch – German/German – English lexicography: its early beginnings. In: Lexicographica. International Annual for Lexicography 1985, pp. 134–164.

Stieler, K. (1968 [1691]): Der Teutschen Sprache Stammbaum und Fortwachs […], Nürnberg: Johann Hoffman. Rpt. with an afterword by Stefan Sonderegger. Munich.

Strickland, L. (2011): Leibniz and the two sophies: the philosophical correspondence. Edited and translated by Lloyd Strickland. Toronto.

Sundby, T./Baruël, J. G. E. (1883–1884): Dansk-Norsk-Fransk Haand-Ordbog 1–2. Copenhagen.

Turton, S. (2019): Unlawful entries: buggery, sodomy, and the construction of sexual normativity in early English dictionaries. In: Dictionaries: Journal of the Dictionary Society of North America 40 (1), pp. 81–112.

Weir, E. (1888): Heath's new German dictionary: in two parts, German-English – English-German. Boston. [Also 1888 as Cassell's New German Dictionary […]. London.]

Woodward, M. E. (1892): A vocabulary of English-Chinyanja and Chinyanja-English as spoken in Likoma. Lake Nyasa […]. London.

Contact information

Nicola McLelland
University of Nottingham
nicola.mclelland@nottingham.ac.uk

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Martina Nied Curcio

DICTIONARIES, FOREIGN LANGUAGE LEARNERS AND TEACHERS

New challenges in the digital era

Abstract In foreign language teaching the use of dictionaries, especially bilingual, has always been related to the hypotheses concerning the relationship between the native language (L1) and second language acquisition method. If the bilingual dictionary was an obvious tool in the grammar-translation method, it was banned from the classroom in the direct, audiolingual and audiovisual methods. Also in the communicative method, foreign language learners are discouraged from using a dictionary. Its use should not obstruct the goals of communicatively oriented foreign language learning – a view still held by many foreign language teachers.

Nevertheless, the reality has been different: Foreign language learners have always used dictionaries, even if they no longer possess a print dictionary and mainly use online resources and applications. Dictionaries and online resources will continue to play an important role in the future. In the Council of Europe's language policy, with its emphasis on multilingualism and lifelong learning, the adequate use of reference tools as a strategic skill is highlighted. In several European countries, educational guidelines refer to the use of dictionaries in the context of media literacy, both in mother tongue and foreign language teaching. Not only is their adequate use important, but so too is the comparison, assessment and evaluation of the information presented, in order to develop Language Awareness and Language Learning Awareness. This is good news. However, does this mean that dictionaries are actually used in class? What role do dictionaries play in foreign language teaching in schools and universities? Are foreign language learners in the digital era really competent users? And how competent are their teachers? Are they familiar with the current (online) dictionary landscape? Can they support their students? After a more in-depth study of the status quo of dictionary use by foreign language learners and teachers and the gap between their needs and the reality, this contribution discusses the challenges facing lexicographers and meta-lexicographers and what educational policy measures are necessary to make their efforts worthwhile in turning foreign language learners – and their teachers – into competent users in a multilingual and digital world.

Keywords Dictionaries; dictionary use; dictionary teaching; dictionary didactics; online resources; foreign language learner; foreign language teacher; language awareness; foreign language teaching; lifelong learning, reference tools, media literacy

1. Dictionaries and lexicographic resources as important reference tools in foreign language learning

In education systems throughout the world, lexicographic products have always been necessary aids to improve language skills and facilitate the study of foreign languages. The significant role that lexicographic activities play in society has been recognised in international politics, inter alia, in 1975 in the Helsinki Final Act of the Conference on Security and Cooperation in Europe.

The classification of lexicography as a cultural practice ("kulturelle Praxis", Wiegand et al. (eds.) 2010 pp. 3, 103) also demonstrates its important pedagogical-cultural role. Although dictionaries have changed in terms of structure, appearance and medium, due especially to globalisation and digitalisation, their importance for society, as well as for the individual, has by no means diminished. On the contrary, dictionaries, encyclopaedias and reference

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

works in general are not only a fundamental tool for translating and learning a foreign language, but their adequate consultation is one of the basic strategies for obtaining new information and accessing the world of general knowledge. The Council of Europe's language policy, with its emphasis on multilingualism and lifelong learning, states that reference works, as well as a high level of research competence and adequate use of strategies are of fundamental importance. (Council of Europe 2001, 2018). Nowadays, with the multiplicity of lexicographic resources, it is particularly important to be familiar with good-quality resources, to have a critical view and to be able to distinguish, with the help of pre-established criteria, what kind of resources are appropriate in a specific situation and context and for a particular task with a precise goal. Well-developed media literacy, with the appropriate use of dictionaries and lexicographic resources, is an essential learning strategy.

2. The use of dictionaries and lexicographic resources in foreign language teaching

In modern language teaching, the use of dictionaries, especially bilingual, was always linked to the hypotheses regarding the relationship between first and second language acquisition and to the associated use of the mother tongue and language comparison as a method and strategy. While the bilingual dictionary was an obvious aid in the grammar-translation method, it was banned from the classroom in the direct, audiolingual and audiovisual methods. In the communicative method, too, learners were not allowed to use a bilingual dictionary under any circumstances, but only if necessary they could use a monolingual one. There was a great fear of reverting to the grammar–translation method. In addition, the opinion that using dictionaries contradicted the goals of communicative foreign language teaching was widespread. Vocabulary acquisition and relatively fast communication competence were not supposed to be clouded by too much reflection on correctness. (Herbst/Klotz 2003, p. 288) Nevertheless, the reality was different: "nearly all students use dictionaries practically every day" (Snell-Hornby 1987, p. 167) This statement is still true today, even if the medium has changed.

As we have already mentioned, the use of dictionaries in foreign language teaching officially became more important again with the publication of the Common European Framework of Reference for Languages (CEFR) (Council of Europe 2001), because the aim of foreign language teaching is not only to improve language competence, but also to successfully cope with foreign language situations. (Herbst/Klotz 2003, p. 288) Learner autonomy, language awareness and the use of strategies became increasingly important and the realisation that dictionaries and their competent use are indispensable for learning a foreign language in the long term was no longer ignored. (Zöfgen 2010, p. 108) As a consequence of this development we can find recommendations on dictionary use in educational guidelines and curricula in various European countries, i.e. Germany and Italy (Nied 2015; Abel in this volume). The use of dictionaries is once again officially allowed; the practice, which had long been common, has thus been legitimized.

Nevertheless, it must be noted that dictionaries and lexicographic online resources and applications are still neglected in foreign language teaching. There are still teachers who are convinced that referring to dictionaries takes too long and interrupts the language learning process, so consulting them on a tablet or smartphone is usually only possible for learners outside class. This means that students are left on their own and are often, therefore, greatly lacking in the knowledge, skills and strategies regarding the use of dictionaries. At the same

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

time, they expect the teacher to present various resources for learning and to explain their use.

Even if there are teachers who want to teach dictionary use, it is up to the individual teacher to decide whether and how much to use a dictionary. Furthermore, they have very little support. The official curricular guidelines are too vague and also exercises in textbooks include tasks such as "work with a dictionary", "use a dictionary" or "the dictionary will help you" – instructions that are actually worthless.

Therefore, it is not clear where and how foreign language learners should learn to use a dictionary. While the use of dictionaries is practised in the teaching of L1 (mainly at primary school level and, unfortunately, often to learn the alphabet or only to look up the meaning of an unfamiliar word (cf. Merten 2011), foreign language teachers often take it for granted that students have learnt to use a dictionary in their mother tongue lessons and can therefore now apply this ability quite naturally to foreign language learning (cf. Bimmel/Van de Veen 2000, p. 38). Foreign language teachers, as we will see in section 3, are often convinced that digital natives are much better at using online dictionaries and apps than they are. This misconception means that no dictionary teaching takes place.

As far as online dictionaries and lexicographic applications are concerned, they are frequently banned from the classroom. Needless to say, this decision is justified if the aim is to assess and/or test a certain language skill, especially vocabulary. However, knowing that students are using online bilingual dictionaries on mobile devices or computers anyway, it would make more sense to show them the better ones. Experimenting with and analyzing different types of dictionaries may help in the language learning process, It would be useful to discuss their pros and cons or problems in their use and, above all, to reflect with the students on their own competence in using them. Moreover, paradoxically, it is often common practice that in official examinations for language certificates only the use of printed dictionaries is allowed.

This misconception about dictionary use has existed for a long time and has meant that there has been very little specific teaching of the topic. As a result, there is no conscious reflection on whether, when and how to use the dictionary. In the age of the Internet, analyzing and discussing online dictionaries and language resources in general - should play a fundamental role in foreign language learning, in order to develop language awareness, language learning awareness and also critical media literacy, as required by the CEFR and consequently by educational guidelines.

3. Foreign language learners as subjects in the Research into dictionary use

Thanks to the emergence of a new field of research, Research into dictionary use, and by developing the theoretical and methodological bases, Herbert Ernst Wiegand (1936–2018) paved the way for empirical studies. This field has gained importance in recent years, particularly since the 1990s, and much empirical research into dictionary use has been carried out by lexicographers and meta-lexicographers. The number of studies has reached such proportions that an overview has become increasingly difficult (cf. Tarp 2009, p. 276). There are now more than 250 empirical studies on dictionary use in the field of foreign language teaching, or with a foreign language learner as user (Nied Curcio 2022). As a result, today, in 2022, one could actually assume that the dictionary user, the former "bekannter Unbe-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

kannter" ('known unknown'), (Wiegand 1977, p. 59) is quite well-known. Unfortunately, this is not necessarily the case because, 1. the classic dictionary has changed fundamentally in its form, 2. foreign language learners rarely use printed dictionaries any more, but rely on online dictionaries and apps, 3. there are more and more hybrid forms, such as dictionaries + grammar tables and dictionaries + text translations, which foreign language learners like to use (Müller-Spitzer et al. 2018, p. 298), and 4. dictionaries are sometimes also completely replaced by automatic translators, Google searches and also extra-lexicographic resources. (Frankenberg-Garcia 2005; Caruso/De Meo 2012; Gromann/Schnitzer 2015)

Another issue must also be addressed: how long the results retain their validity. Many studies, in which the printed dictionary was the subject, may have lost their 'eternal validity' due to digitalization and today's almost exclusive use of electronic and online dictionaries. Although the results are still interesting, they need to be re-analysed and re-examined along with dictionaries currently in use. The act of usage takes place in a totally different way and the difficulties or even errors in using these resources are different from those of using a printed dictionary or an electronic one in the 1990s. We should also consider whether the situation and context of usage and usage task would still be authentic in today's teaching.

For this reason, research into dictionary use is, in my opinion, in a very challenging situation, precisely because the object of study is online dictionaries, applications, hybrid forms etc. and all these are constantly evolving (and improving). This also applies to translation programmes such as Google translator or DeepL. Changes and updates of online dictionaries and translation programmes often take place without the user being informed or aware of them. They are usually corpus-based and algorithmic resources and it is a constantly evolving process. If studies were carried out today, the results could lose their validity after a short space of time. This was not the case when the printed dictionary was the object of study, because often years passed between one edition and another, and the differences between the two editions could also be studied. Dictionary criticism and research into dictionary use could influence the lexicographic process and new editions, which is almost impossible today.

3.1 Foreign language learners and their dictionary use competence

It is mainly to the credit of Andreas Herbert Welker that an overview was presented for the first time (2006, 2010). In his *overview* Welker (2010) proposes a division into six categories: 1. surveys, 2. studies on actual dictionary use, 3. studies on the effect of dictionary use, 4. studies on specific dictionary features and on specific dictionaries, 5. research on the use of electronic dictionaries and 6. research on the teaching of dictionary use. In recent decades, a number of empirical studies on the use of electronic and online dictionaries and also on the use of online resources in general during the learning process have been published. The survey, and especially the questionnaire, is still one of the most applied methods, while studies on paper dictionaries are, for obvious reasons, disappearing. In addition, it can be observed that in recent years there has been an increase in the number of studies carried out in a concrete situation, with a specific task and with the aim of obtaining information on the effect of use. There are still relatively few studies on the effect of dictionary didactics.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

In the following chapters, the results of studies spanning some 40 years (1979–2021) are presented in an extremely concise way.[1] Most of the empirical research concerns the use of dictionaries in the context of English as a foreign language by learners of different L1 languages, e.g. Arabic, Chinese, Japanese, Swedish or groups of learners with several L1 languages. There are far fewer studies on the use of dictionaries for other foreign languages. However, as can be seen, the foreign language learner's behavior in using the lexicographic resource depends little on L1 and L2. In the description we follow Welker's categorization, omitting 4. because it is less relevant to the field of foreign language teaching.

3.1.1 Surveys

The following results emerge from the questionnaire studies of recent years[2]:

- Not surprisingly, foreign language learners prefer the bilingual dictionary.
- The monolingual dictionary is mainly used at an advanced level of proficiency. It is assumed that this is as a result of the teaching method and/or the advice given by the teacher.
- It is also interesting to note that dictionary use generally decreases as the learner reaches a higher language level, i.e. when he/she acquires advanced competence.
- Regarding the situation and context of usage, it has been ascertained that dictionaries are mostly used in translation, written reception and written production.
- Also in relation to the situation and context of usage, foreign language learners mainly look for the meaning of an unknown word as part of the decoding process. They often also look for pronunciation and grammatical information.
- In bilingual dictionaries, students concentrate on finding equivalents.
- Foreign language learners go directly to the information they are looking for and do not read the whole dictionary entry. Moreover, most students do not read the introductory notes (preface, instructions for use) before using it. The most important thing for students is that the search leads quickly and directly to a result.
- Many students are dissatisfied with dictionaries due to a) the lack of the headword, b) the definition and/or explanation, c) the examples.
- They also complain that the entries and explanations (especially in monolingual dictionaries) are too long and/or complex.
- With regard to the bilingual dictionary, students are not satisfied because they are confused by the large number of equivalents and have difficulty in choosing the appropriate one for a specific context of usage.

Of course, dictionaries are not always satisfactory. They are not complete, have gaps, are complex and not always easily accessible or user-friendly. However, not all mistakes in using a dictionary are due solely to the dictionary and its content. Several studies indicate that users are not able to use a dictionary adequately. There are various reasons for this. Foreign language learners are not familiar with the overview of dictionaries and do not know which

[1] For obvious reasons, the individual studies cannot be listed by name.

[2] These are results that are repeated in very many studies and can therefore be listed as frequent. Due to the varying number of studies in terms of results, percentages are avoided.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

are suitable for their needs, so they rarely use different types of dictionaries. Indeed, many beginner and intermediate learners are not familiar with their dictionaries and often have unreasonable demands on them. What also often happens is that students do not notice metalinguistic/grammatical information – such as indications on the gender or the regency of verbs (valency) – within a lexicographic entry, or have difficulty deciphering abbreviations and symbols.

These results are certainly also due to the fact that the majority of foreign language learners have not previously received instruction or training in the use of dictionaries, while learners with dictionary training are definitely more experienced because they are skilled users (cfr. 3.2.5).

3.1.2 Studies on actual dictionary use

For many years, researchers in the field of dictionary use did not focus on the user in actu, i.e. the user who, at the moment of the research, is in a concrete and authentic situation of usage, but on potential users, ex actu or post actum. Instead, in order to examine how dictionaries are 'really' used, it is necessary to observe the user in actu (preferably with external observers).

In this chapter, studies that focus on the user in actu are summarized. The most applied methods are observations (including video recordings), usage records, experiments and analysis with log files and eye-tracking. Written usage records are the most applied method. With the increasingly frequent use of electronic dictionaries, studies using log files have also increased. The use of think-aloud-protocols and eye-tracking is still rather rare.

In relation to the results described in this paragraph, it must be considered that the majority of studies concern translation exercises into both L1 and L2, even within foreign language teaching. Translation as a task is employed in the grammar-translation method. We can observe that the results are very similar to those obtained through the questionnaires:

- Dictionaries are mostly used in the decoding process and/or during translation.
- If subjects are free to choose their own dictionary, they prefer a bilingual one.
- In bilingual dictionaries, learners focus on searching for equivalents and/or example sentences.
- A monolingual dictionary is used if the bilingual dictionary does not provide sufficient contextual information about a word.
- Participants usually scroll through the various meanings within an entry rather quickly, focusing mainly on information indicating the meaning, until they are convinced they have found the right meaning. At this point, they proceed to read more carefully.
- Other researchers have found that users normally read the first definition, but often do not even look at the second one or do not complete the reading of the entry.
- In addition, detailed eye-tracking data found that users generally read dictionary entries from top to bottom rather than from bottom to top.
- There is a tendency for foreign language learners to search for single words and much less for phrasemes or parts of sentences.
- Studies have shown that there is a correlation between language level and adequate dictionary use.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

- Many studies with the user in actu have revealed difficulties in the use of dictionaries which are almost identical to those found through questionnaires, for example:
- When translating into the foreign language, many students have difficulties in selecting the appropriate equivalent for the specific context.
- Some problems occur especially with common language words, polysemous verbs, homonyms, phrasal verbs and phrasemes.
- Another difficulty mentioned by the students is the fact that the dictionary entry (especially the monolingual one) is too long and complex.
- In some studies, the difficulties lie in the fact that the information sought is missing or the users are unable to find it.

A number of scholars argue that difficulties in the use of dictionaries result from the fact that students do not have adequate competence in dictionary use. The students ignore metalinguistic/grammatical indications and do not use hyperlinks in online dictionaries, or do not use a dictionary at all. In the case of lexical gaps students prefer to ask the teacher instead of using a dictionary.

In contrast to this, successful and satisfied users use more resources. In general, they are students with advanced linguistic competence.

As already mentioned above, these results are very similar to those of the questionnaires, but less homogeneous as partly contradictory results also emerge. The reasons for this are various: the different design of the research, the selection of subjects, the varying number of subjects, the different tasks and their degree of difficulty, the duration of the research, the experience in using dictionaries and, not to be forgotten, the language level of the foreign language learners, the mother tongue.

3.1.3 Studies on the effect of dictionary use

The most frequent tasks to measure the effect of dictionary use are reading (written reception), writing (written production) and translation. Often, when researchers use the word writing, they mostly mean writing sentences and not texts. In general, subjects are asked to write single sentences. The task of translating is similar, as subjects are mostly not required to translate whole texts, but only isolated sentences or words extracted from texts. In reading tests, users sometimes do not read the texts, but are asked to insert words in the empty spaces within isolated sentences, or to translate words or sentences without differentiating whether the reported difficulties were in understanding or in not finding a correct equivalent. In a few studies, subjects are asked to correct sentences in a foreign language that contained errors typical for this type of learner.

Another point of discussion regarding this type of study is that users are often only given excerpts or single entries from dictionaries, i.e. they do not really have a dictionary to hand. A further problem is that the results of the studies are often linked to specific products, so that comparison between studies is difficult and it can be argued that for this very reason there is no generally valid statement.

There are also studies that have found no significant difference in effectiveness between dictionary use and non-use and others that state the exact opposite. The same applies to the difference between bilingual and monolingual dictionaries, and also between printed and electronic ones.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Furthermore, some scholars believe that consulting a dictionary is of no help when it comes to memorizing new words. Wolfer et al. (2016) argue that dictionary use can be useful but only in the case where the user, in the first instance, realises that he or she is faced with a language problem to be solved and then, in the second instance, uses the lexicographic resource. This relationship between language awareness and the (adequate) use of a dictionary is also discussed in some studies which had not focused on this aspect in their design and research objective (Müller-Spitzer et al. 2018). We have to say that there are few 'real' studies on the effectiveness of the dictionary. Moreover, their results are very inconsistent and often contradictory.

3.1.4 Research on the use of electronic dictionaries and online resources

With the arrival of electronic dictionaries in the 1990s, studies initially focused on their use as opposed to the use of printed dictionaries. From the very beginning, foreign language learners appeared motivated to use electronic dictionaries and this has been confirmed recently. Today, we know that students mainly use online dictionaries and applications, also on mobile devices, to overcome language difficulties. The use of smartphones offers foreign language learners an almost unlimited choice of possibilities to overcome existing language difficulties in a matter of seconds and mostly free of charge. We know that many foreign language learners no longer buy a printed dictionary and do not spend money on online access. They appreciate the fast, easy access of online dictionaries and the fact that they are free and always up-to-date. They also like the fact that by entering the first letters in the search engine, they are guided to the respective entry and they find the spell-checker very useful. However, it is precisely this speed that leads to the inappropriate use of the resource (Müller-Spitzer et al. 2018).

In my opinion, when comparing the use of printed dictionaries with the use of electronic ones, it is striking that there are parallels in user behaviour. The same difficulties emerge: disorientation, lack of knowledge about dictionaries, looking up single words, choosing the first equivalent, mainly consulting examples etc. Students generally do not read all the information in the entry and do not 'scroll down' but tend to focus on the part that is directly visible on the monitor. Furthermore, detailed eye-tracking data reveals that users generally proceed from the top of the entry downwards rather than from the bottom upwards. This vertical reading seems to be one of the reasons why students do not see the solutions offered in bilingual dictionaries on the right side of the dictionary entry and consequently the consultation is not successful (Nied Curcio 2014; Runte 2015; Müllers-Spitzer et al. 2018). So far, we can only speculate on the causes: either the electronic version is identical to the printed version, or the students use the online dictionary in the same way as a printed dictionary. From this point of view, the negative results seem even more serious, as the technical potential actually offers unlimited search possibilities.

It seems that the behavior of language learners is also changing. Recent studies have shown that more and more learners are looking up words in a search engine. Search engines seem to be taking over the main functions of a monolingual dictionary, such as providing definitions or examples, and partially replacing bilingual dictionaries, providing equivalents and spelling. The act of consulting an online dictionary also increasingly resembles the use of a search engine because students expect the online dictionary to 'behave' like a search engine. In search engines, users often enter the unknown foreign word, together with a metalinguistic term or with a kind of key word, e.g.: "Konjunktiv 2 mit wenn" or "deshalb significado" (Müller-Spitzer et al. 2018, p. 292). It is also interesting that, in the same study, students

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

distinguish between dictionaries and automatic translators. In dictionaries, they tend to look for single words, whereas in automatic translators they enter complex words, syntagmas and complete sentences.

To solve problems in the foreign language, today's foreign language learners do not only use lexicographic online resources, but also extra-lexicographic resources and combinations of resources that also include dictionaries with grammar tables, dictionaries with automatic translators and dictionaries that are based on parallel texts.

Furthermore, it could be stated, that there is a correlation between the language level (also of L1), the language awareness, the ability to use strategies and an adequate use of dictionaries. The higher the linguistic level, the more the students are able to use strategies; the longer they reflect, the more satisfactory and adequate the use of the online resource is. The level of language awareness seems to be the crucial prerequisite for the competent use of dictionaries and lexicographic online resources (Frankenberg-Garcia 2011; Nied Curcio 2020).

3.1.5 The effect of dictionary teaching

There are very few empirical studies on the effectiveness of dictionary teaching, but they show a significant improvement in the use of dictionaries by skilled users, who also improved their search strategies and were able to reduce errors in the foreign language (Lew/Galas 2008; Welker 2010, pp. 313–321). Targeted teaching of word combinations, e. g. collocations, phrasal verbs and idioms, and looking them up in dictionaries meant that students' attitudes improved and errors in this field decreased. Students learned that words have relationships with each other and how important it is to look up combinations of words as well as a single word.

When students are asked, they express a desire to learn more about the lexicographic tools available. They want to know which language learners' dictionaries are available, which are the most valid, how they are designed and structured and how to recognize reliable information. They show great interest in improving their skills in using dictionaries and online resources, with the aim of making fewer mistakes in the foreign language.

4. Foreign language teachers and their dictionary use competence

Studies on foreign language teachers' competence in dictionary use are almost non-existent. Based on my experiences in training courses for foreign language teachers, it seems that teachers are still experienced users of printed dictionaries, but are not very familiar with online dictionaries and applications. As we have already mentioned, the use of bilingual dictionaries is often not allowed in the classroom and online dictionaries and applications are almost completely excluded.

In order to learn more about teachers' competence in using online resources and apps, I carried out a small research project during a workshop on dictionary teaching[3] in 2017, with 50 teachers of German as L2 in Italy, using a multi-methodological approach. A question-

[3] From here on, the term dictionary teaching is used in a broader sense. It refers not only to dictionaries, but also to lexicographic online resources, hybrid forms (i.e. dictionaries and grammar tables) (cf. 3.), search machines and translation programmes.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

naire about the teachers' use of lexicographic tools in general and in class was distributed at the beginning of the teacher training course. The main part of the course was a workshop where the teachers were asked to experiment freely with various online dictionaries and apps and also translation programmes. There were breaks for discussion and reflection on the tools and the teachers' own user behaviour. At the end of the course the teachers completed another questionnaire in order to assess whether the workshop had been useful and, more importantly, whether their attitude towards the use of online dictionaries and apps in class had changed. Some of the main results will be listed below:

- The first questionnaire, comprising general questions, confirmed that the majority of teachers allow the use of a printed dictionary in class (46 subjects give their students permission to use a bilingual printed dictionary and 5 allow the use of a monolingual printed dictionary), but only 18 teachers permit the use of online dictionaries on tablets and smartphones and only 7 of them allow the use of applications in class.
- The four teachers who do not allow the use of dictionaries in their classroom justify their decision with the following reasons:
    - students will use words they know,
    - the language level is too low,
    - students have difficulties in using dictionaries,
    - students are too distracted,
    - the use of smartphones is forbidden,
    - no computers/tablets are available
- Instead, almost all the teachers (49) allow the use of dictionaries for homework.
- More than half of the teachers (27) think that students are able to use online dictionaries and related applications; 22 teachers are of the opposite opinion and one teacher did not answer.
- The results show that 16 teachers use monolingual printed dictionaries and the same number of teachers also use bilingual printed dictionaries.
- 21 teachers use online dictionaries on their smartphones, and 8 use apps on smartphones and tablets.
- What is striking is the teachers' self-assessment data. 34 teachers admit that they are not familiar with the use of online dictionaries and lexicographic applications.
- When asked if they also use translation programs, only 4 teachers reported using them.[4]

The second questionnaire was completed after the dictionary training session. Due to the short duration of the workshop, it is unreasonable to expect that the teachers could have become fully-informed and skilled users. All teachers (50) indicated that the course had been very useful, that they had enjoyed it very much and that they especially felt that they were now more familiar with online dictionaries and lexicographic applications. The aspect of learning by doing, of exploring and experimenting with the various resources at first hand, and of comparing and evaluating them was rated as very positive. 22 teachers appreciated the fact that they had discovered many new online dictionaries and applications, and had thus gained a better overview of existing resources that can be used for teaching. They

[4] This could also be the problem of desirability: "Are subjects saying [...] what they do, or what they think they ought to do, or indeed a mixture of all three?" (Hatherall 1984, p. 184) .

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

felt that it was useful to learn more about the structure of a dictionary and the microstructure of a lexicographic entry and to understand how translation programmes work.

It is therefore clear that dictionary teaching can have a beneficial effect and can even influence or change the participants' attitude. Some teachers decided to start using online dictionaries or to pay for an online dictionary or to reflect on how to incorporate the use of online dictionaries and applications better in their teaching. When asked whether they would now, after the course, incorporate online dictionaries and apps into their German lessons, all 50 teachers said yes. This shows that even the most skeptical, and those who had indicated that they would not allow the use of online dictionaries in class, had changed their minds. Nearly all teachers (47) would be willing to participate in a further course and would especially like to learn specifically how to include the use of online dictionaries and applications in their teaching. This paves the way for facing new challenges.

5. New challenges in the digital era: lexicographic tools in foreign language teaching

As we have seen, printed dictionaries are rapidly disappearing from the daily lives of foreign language learners and at this point it is not yet clear what role dictionaries will actually play not only in future foreign language teaching, but also in the area of academic and specialist knowledge acquisition. The dictionary as a lexicographic reference work, in its printed form, was a very specific physical object. With increasing digitalization, not only has its structure changed, but as we have already mentioned, hybrid forms have emerged, such as dictionary + grammar table, dictionary + grammar table + parallel text or even dictionary + parallel text + automatic translator, resources which were physically separate before the digital age. In addition, the overwhelming number of resources means an overall view is impossible. Often, neither the function nor the purpose of the resources offered is clear, and the authors are generally not explicitly mentioned. The quality is therefore no longer transparent. Many resources are updated automatically and constantly, so that the individual stages of the update are no longer distinguishable. This profound change in lexicographic practice is very often not perceived by foreign language learners, even though they regularly use these reference works. The data from research into dictionary use show how much lexicographic resources have changed in recent decades, and more strikingly, that there are parallels between printed dictionaries and online dictionaries in terms of usage behavior and users' difficulties over the same period. Perhaps this behavior is gradually changing and converging with the use of search engines and translation tools. This has not yet been confirmed from research, but it means that empirical studies have to be carried out in this direction.

If we focus on the foreign language learner as user in lexicographic practice, the potential user and the addressee must be linked together, which means that the development of a lexicographic resource should realistically be conceived with the potential user in mind. Very often, too many addressees are mentioned in dictionaries (also for commercial reasons). Lexicographic practice (and also theoretical discourse) should carefully consider the results of the research into dictionary use and thus focus more on the potential user and, in our case, on the foreign language learners' profile. For example, it would be extremely useful to create a specific online portal for a specific language learner profile, with the various dictionaries suitable for this type of user (similar to the one for linguistics students (<http://www.linse.uni-due.de/>).This portal would provide helpful information i.e.: a) the macro-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

and micro-structure of the dictionaries presented could be explained and commented on and there could be suggestions on how to use them (dictionary criticism at the service of the user); b) at the same time these listed resources could have hyperlinks. Existing portals are usually uncommented lists of dictionary titles. The selection criteria are not clear, but it is certainly not oriented towards the profile of a specific user. Such a portal for students and teachers, of a specific foreign language, should be conceived and created in collaboration with renowned lexicographic institutions and lexicographers, in cooperation with institutions responsible for education and teacher training, and perhaps even with the cooperation of teachers and their students.

User-orientation is also extremely important for future studies in the field of research into dictionary use, although the validity of the results could sometimes quickly become outdated due to the constant updating of online resources. The studies and results mentioned above invite us to undertake more extensive research, not limited exclusively to lexicographic resources, but also including search engines and automatic translators; they also urge us to investigate research competence and media literacy in general. In a sense, despite all the research undertaken in recent decades, the change in the "dictionary" medium has meant that we are once again faced with a "known unknown". Research into dictionary use can indeed build on previous studies, but it must focus on this 'new' user, the digital native, whose approach to the use of online lexicographic tools on the one hand seems to be identical to the use of a printed dictionary, but on the other hand is also changing and moving towards the use of a search engine. Subsequently, it is extremely important to discuss the consequences for lexicographic practice.

At the same time, it is also necessary to work at an educational policy level. In the digital, global and multilingual world, which is characterized by lifelong learning, well-developed media literacy (with the appropriate use of dictionaries and online lexicographic resources) is essential as a learning strategy. Unfortunately, foreign language teaching has not really realized these new requirements, even though the CEFR explicitly refers to the importance of this competence and despite the fact that many European countries' educational guidelines include the use of dictionaries/online resources explicitly in foreign language teaching. Online lexicographic resources should no longer be banned from foreign language teaching. It is neither sufficient to criticise the dictionaries, resources and online applications that learners use, nor to leave students on their own. After all, foreign language learners use dictionaries, with or without training. It would be much better to integrate the research tools they use into lessons, to reflect on their use and to enable learners to become experienced and skilled users.

There is a vital need for teaching how to use modern lexicography resources, and foreign language courses could be an excellent place in which to do this. However, teachers are not necessarily competent users of online lexicographic resources. Consequently, dictionary teaching cannot be implemented directly, but teachers should be trained first.

Efforts should also be made on another front: to set up a forum with publishers and textbook authors to focus on concrete exercises designed for various language activities that are clearly defined in their objective.

The biggest challenge is likely to be in bringing together the different fields of research into dictionary use, foreign language acquisition research, foreign language teaching and didactics, teacher training and lexicographic practice. In concrete terms, this will mean professionals from the various disciplines collaborating creatively with the aim of enabling for-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

eign language learners to become skilled, and successful users of online dictionaries and lexicographic resources and, in a broader sense, autonomous users in terms of critical media literacy. Above all, the goals should be for the foreign language learner to become a more 'known' user again and to respond more adequately to the digital user's needs in the various fields. Lexicographic resources should once again become a useful tool for foreign language users and their learning process in this Third Millennium.

References

Bimmel, P./Ven, van de M. (2000): Man nehme ein Wörterbuch… In: Fremdsprache Deutsch. (Themenheft: Übersetzen im Deutschunterricht) 23, pp. 38–39.

Caruso, V./De Meo, A. (2012): What else can databases do to assist translators? Illustrating a rated inventory of Web dictionaries. https://www.researchgate.net/publication/235727124 (last access: 15-04-022).

Council of Europe (2001): Common European Framework of Reference for Languages: learning, teaching, assessment. Cambridge.

Council of Europe (2018): Common Europe Framework of Reference for Languages: learning, teaching, assessment. Companion volume with new descriptors. Strasbourg. https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989 (last access: 18-04-2022).

Frankenberg-Garcia, A. (2005): A Peek into What Today's Language Learners as Researchers Actually Do. In: International Journal of Lexicography (IJL) 18 (3), pp. 335–355.

Frankenberg-Garcia, A. (2011): Beyond L1-L2 equivalents: where do users of English as a foreign language turn for help? In: International Journal of Lexicography (IJL) 24 (1), pp. 97–123.

Gromann, D./Schnitzer, J. (2015): Where do business students turn for help? An empirical study on dictionary use in foreign-language learning. In: International Journal of Lexicography (IJL) 29 (1), pp. 55–99.

Hatherall, G. (1984): Studying dictionary use: some findings and proposals. In: Hartmann, R. R. K. (ed.): LEXeter '83 Proceedings. Tübingen, pp. 183–189.

Herbst, Th./Klotz, M. (2003): Lexikographie. Paderborn/München.

Lew, R./Galas, K. (2008): Can dictionary skills be taught? The effectiveness of lexicographic training for primary-school-level Polish learners of English. In: Bernal, E./De Cesaris, J. (eds.): Proceedings of the XIII EURALEX International Congress: Barcelona, 15–19 July 2008. Barcelona, pp. 1273–1285.

Merten, S. (2011): Arbeit mit Wörterbüchern. In: Pohl, I./Ulrich, W. (eds.): Wortschatzarbeit. Baltmansweiler, pp. 348–360.

Müller-Spitzer, C./Domínguez Vazquez, M. J./Nied Curcio, M./Silva Dias, I. M./Wolfer, S. (2018): Correct hypotheses and careful reading are essential: results of an observational study on learners using online language resources. In: Lexikos 28, pp. 287–315.

Nied Curcio, M. (2014): Die Benutzung von Smartphones im Fremdsprachenerwerb und -unterricht. In: Abel, A./Vettori, Ch./Ralli, N. (eds.): Proceedings of the XVI EURALEX International Congress: The User in Focus. 15–19 July 2014. Bolzano/Bozen, pp. 263–280.

Nied Curcio, M. (2015): Spielen Wörterbücher bei der Sprachmittlung noch eine Rolle? In: Nied Curcio, M./Katelhön, P./Basic, I. (2015): Sprachmittlung – Mediation – Mediazione linguistica. Ein deutsch-italienischer Dialog. Berlin, pp. 291–317.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Nied Curcio, M. (2020): Sprachbewusstheit als wichtige Voraussetzung bei der Recherche in mehrsprachigen Online-Ressourcen. In: Hepp, M./Salzmann, K. (eds.): Sprachvergleich in der mehrsprachig orientierten DaF-Didaktik. Theorie und Praxis. Rome, pp. 85–109.

Nied Curcio, M. (2022): L'uso del dizionario nell'insegnamento delle lingue straniere. Rome.

Runte, M. (2015): Lernerlexikographie und Wortschatzerwerb. Berlin/Boston: de Gruyter.

Snell-Hornby, M. (1987): Towards a learner's bilingual dictionary. In: Cowie, A. P. (ed.): The Dictionary and the Language Learner. Papers from the EURALEX Seminar at the University of Leeds, 1–3 April 1985. Tübingen, pp. 159–170.

Tarp, S. (2009): Reflections on lexicographical user research. In: Lexikos 19, pp. 275–296.

Welker, H. A. (2006): O uso de dicionários: Panorama geral das pesquisas empíricas. Brasilia.

Welker, H. A. (2010): Dictionary use: a general survey of empirical studies. Brasilia.

Wiegand, H. E. (1977): Nachdenken über Wörterbücher: Aktuelle Probleme. In: Drosdowski, G./Henne, H./Wiegand, H. E. (eds.): Nachdenken über Wörterbücher. Mannheim/Wien/Zürich, pp. 51–102.

Wiegand, H. E./Beißwenger, M./Gouws, R. H./Kammerer, M./Mann, M./Storrer, A./Wolski, W. (eds.) (2010): Wörterbuch zur Lexikographie und Wörterbuchforschung. Dictionary of lexicography and dictionary research. Vol. 1 (A–C). Berlin/Boston.

Wolfer, S./Bartz, Th./Weber, T./Abel, A./Meyer Ch. M./Müller-Spitzer, C./Storrer, A. (2016): The effectiveness of lexicographic tools for optimising written L1-Texts. In: International Journal of Lexicography (IJL) 31 (1), pp. 1–28.

Zöfgen, E. (2010): Wörterbuchdidaktik. In: Königs, F. G./Hallet, W. (eds.): Handbuch Fremdsprachenunterricht. Seelze/Velber, pp. 107–110.

Contact information

Martina Nied Curcio
Università degli Studi Roma Tre
martina.nied@uniroma3.it

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Part III: Proceedings of Talks, Posters, and Software Demonstrations

Dictionaries and Society

Stefan Engelberg

LEXICOGRAPHY'S ENTANGLEMENT WITH COLONIALISM: THE HISTORY OF TOK PISIN LEXICOGRAPHY AS COLONIAL HISTORY

Abstract Tok Pisin is a pidgin/creole language spoken since the late 19th century in most of the area that nowadays constitutes Papua New Guinea where it emerged under German colonial rule. Unusual for a pidgin/creole, Tok Pisin is characterized by a extensive lexicographic history. The Tok Pisin Dictionary Collection at the Leibniz Institute for the German Language, described in this article, includes about fifty dictionaries. The collection forms the basis for the sketch of the history of Tok Pisin lexicography as part of colonial history presented here. The basic thesis is that in the history of Tok Pisin, lexicographic strategies, dictionary structures, and publication patterns reflect the interest (and disinterest) of various groups of colonial actors. Among these colonial actors, European scientists, Catholic missionaries, and the Australian and US militaries played important roles.

Keywords Pidgin; Tok Pisin; colonialism; history of lexicography; lexicography and war; missionary linguistics; colonial linguistics

1. Tok Pisin and the colonial history of New Guinea

Tok Pisin is a pidgin/creole language spoken in eastern New Guinea, the Bismarck Archipelago and the northern Solomon Islands. It is one of three closely related pidgin-creoles beside Bislama, spoken in Vanuatu, and Pijin in the Solomon Islands. Tok Pisin has its roots in English-based pidgin varieties of the southwestern Pacific and formed in the late 19th century under German colonial rule. Its origins are closely connected to the plantation economy in colonial Oceania and the – often forced – labor migration in the South Pacific on which this economy was based. The canefield plantations in Queensland and the mostly German-owned coconut plantations in Samoa with the pidgins developing there played a special role in this process. Workers returning from there to the Bismarck Archipelago formed the nucleus for the development and spread of Tok Pisin.[1]

New Guinea has long been inhabited by many independent and mostly very small groups of people speaking more than a thousand languages. It was not until the 19th century that European imperial powers began take an interest in New Guinea. The western part of the island was claimed by the Netherlands in 1828 and, in 1963, became part of Indonesia under circumstances contrary to international law. In 1884, Germany claimed the northeastern part of New Guinea, the Bismarck Archipelago, and the Northern Solomon Islands, and Britain took possession of southeastern New Guinea. German rule, to which also the considerable lexical influence of German on early Tok Pisin can be attributed, ended effectively in 1914 and officially in 1919. After WW I, both eastern territories of New Guinea came under Australian administration. With a brief interruption due to the Japanese occupation in WW II, Australian rule lasted until 1975, when Papua New Guinea gained independence.

[1] For the complex prehistory of Tok Pisin cf., e. g., Mühlhäusler (1978, 1979) and Baker (1993).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Tok Pisin is nowadays one of the official languages of Papua New Guinea along with English and Hiri Motu. The majority of the population in Papua New Guinea uses it as a second language, and the number of L1 speakers is growing.

2. The Tok Pisin Dictionary Collection at the IDS

As with many pidgin languages, the bad reputation Tok Pisin had among Europeans and Australians (cf. Engelberg 2014) stood in stark contrast to its indispensability for the colonial economy, missionary work, and the exercise of power. This partly explains the extensive lexicographic history of Tok Pisin with several dozen quite comprehensive dictionaries and numerous smaller vocabularies.

Copies of about fifty Tok Pisin dictionaries have been collected at the Leibniz Institute for the German Language (Leibniz-Institut für Deutsche Sprache, IDS) where German language contact in the former German colonies has been studied (Engelberg/Stolberg 2017). The following list documents the content of this collection in chronological order.

N°	Year	Reference	Comment
01	1902 / 1913 [ca.]	Dempwolff, Otto ([ca.] 1902/1913): Pidgin-Englisch von Deutsch-Neuguinea. Wörterverzeichnis. – Manuscript. Archive: Universitätsarchiv Hamburg | Best. 305h: Fachbereich Asien-Afrika-Wissenschaften (Asien-Afrika-Institut), Nr. 206.	*TP > German; undated; prob. from the very early 20th c.; marginalia prob. from 1913.*
02	1913	Thurnwald, Richard (1913): Ethno-psychologische Studien an Südseevölkern auf dem Bismarck-Archipel und den Salomo-Inseln. – Leipzig: Verlag von Johann Ambrosius Barth.	*Including: word list TP > German; TP > Buin > German.*
03	1913 [ca.]	Dempwolff, Otto ([ca.]1913): Pijin. Wörterverzeichnis. [fragment A-B]. – Manuscript. Archive: Universitätsarchiv Hamburg | Best. 305h: Fachbereich Asien-Afrika-Wissenschaften (Asien-Afrika-Institut), Nr. 206.	*TP > German; undated; prob. from Dempwolff's stay in New G. in 1913.*
04	1924	Brenninkmeyer, [Pater] Leo (1924): Einfuehrung ins Pidgin-Englisch. Ein Versuch. – Typoscript, mimeographed. Kamanacham [New Britain, PNG].[2]	*Includes short thematic word lists German > TP.*
05	1926	Borchardt, Karl (1926a): Tok Boi Wörterbuch. – Typoscript, mimeographed. [Manus, Admirality Islands, PNG].	*TP > German / English.*
06	1926	Borchardt, Karl (1926b [ca.]): Kleines Woerterbuch / Deutsch – Tokboi. – Typoscript, mimeographed. [Manus, Admirality Islands, PNG].	*German > TP.*
07	1929	Blackwood, Beatrice (1929): Pidgin & English-Petets Vocabulary. Rabaul [New Britiain, PNG]. – Archive: Alexander Turnbull Library / National Library of New Zealand | Project: Miscellaneous Series microfilm | North Solomon Islands - Language | Micro-MSColl-20-2814 [=Beatrice Blackwood Papers / Pitt-Rivers Museum, Oxford University, Parks Road, Oxford, OX1 3PP / The National Library of Australia, State Library of New South Wales 1988 / Reel 9 / A. Working Papers / Box 7: N. Solomons. 1929-1930: Language.	*Dictionary English > Petets, supplemented by entries Englisch > TP.*

[2] A typewritten dictionary by Brenninkmeyer from 1925 is still missing from our collection.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Nº	Year	Reference	Comment
08	1930	van Baar, William ([ca.]1930): Pitshen-Wörterbuch. – Typoscript, mimeographed. Mugil [New Guinea, PNG].	*German > TP.*
09	1935	Anonymous (1935): Pijin Lexikon. – Typoscript, mimeographed. Alexishafen [New Guinea, PNG].	*TP > German; also as „Wörterbuch mit Redewendungen".*
10	1937	Haslett, E . [„Maski Mike"] (1937): Pidgin English Dictionary of Common Nouns and Phrases used in Conversation with Natives in the Territory of New Guinea. – Townsville: T. Willmett & Sons (Pty.).	*TP > English.*
11	1940 [ca.]	Kutscher, [P.] ([ca.] 1940): Wörterbuch deutsch-pidgin-english. – Typoscript, mimeogr. Vunapope [New Britian, PNG].	*German > TP.*
12	1941	Command of the Military Board (1941): Handbook of Pidgin English. Aboriginal and South Sea Islands. – Melbourne: Military Board, Army Headquarters.	*English > TP.*
13	1943	Hall, Robert Anderson Jr. (1943): Melanesian Pidgin English. Grammar, texts, vocabulary. – (Special Publications of the Linguistic Society of America.) Baltimore, MD: Waverly Press. [„Identical with the Edition published for the United States Armed Forces Institute, Madison, Wisconsin."]	*TP > English; Tok Pisin lemmas rendered in phonetic spelling.*
14	1943	Helton, E. C. N. (1943): Booklet on Pidgin English as used in the mandated territory of New Guinea. With Dictionary of Nouns and Phrases. – Brisbane: W. H. Adams.	*TP > English.*
15	1943	Sayer, Edgar Sheappard (1943): Pidgin English. A Text Book, History, and Vocabulary of Pidgin English, for Writers, Travellers, Students of the English Language and Philologists. – 2. ed. Toronto: E. S. Sayer [author's edition].	*English > pidgin; unspecific mixture of several pidgins.*
16	1944	Army Education Branch, Morale Services Division, Army Service Forces (1944): Melanesian Pidgin English Language Guide. First Level. – Washington, DC: United States Government Printing Office.	*English > TP; thematically organized word lists.*
17	1945	Schebesta, [Rev. Father] Josef & [Rev. Father] Leo Meiser (1945): Dictionary of "Bisinis English" (Pidgin-English). – Revised by Leo Meiser. Typoscript, mimeographed. Alexishafen [New Guinea, PNG].	*TP > English.*
18	1949 / 1957	Dahmen, Johannes (1949/1957): Pidgin-English Dictionary. – Typoscript, mimeographed. Bundralis [Admirality Islands, PNG] / Rabaul [New Britain, PNG].	*TP > English.*
19	1950-1960	Smythe, W. E. ([ca.] 1950-1960): Pidgin Vocabulary. – Manuscript. [Manus, Admirality Islands, PNG].	*TP > English.*

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Nº	Year	Reference	Comment
20	1955-1960 [ca.]	Jäschke Ernst ([ca.] 1955-1960): Wörterkartei Deutsch – Pidgin [G-Z] – Archive: Landeskirchliches Archiv der Evangelisch-Lutherischen Kirche in Bayern | 6.7.0003 Mission EineWelt (MEW) | TB 6 Sprachwissenschaftliches und Ethnologisches | 1.2.15 Pidgin - Tok Pisin | [164] Jäschke, Ernst, Wörterkartei Deutsch - Pidgin - G-H und I-K, Signatur: 6.53./1 / [165] Jäschke, Ernst, Wörterkartei Deutsch - Pidgin - L - N; Signatur: 6.53./2 / [166] Jäschke, Ernst, Wörterverzeichnis Karten Deutsch - Pidgin - O-Z, Signatur: 6.164.	*German > TP, probably from the late 1950s; card index box A - F missing.*
21	1957	Mihalic, Francis (1957): Grammar and Dictionary of Neo-Melanesian. – Techny (Illinois): The Mission Press.	*TP > English; English > TP.*
22	1966	Murphy, John J. (1966): The Book of Pidgin English. Revised edition. – Brisbane: W. R. Smith & Paterson.	*TP > English; English > TP.*
23	1969	Balint, Andras (1969): English – Pidgin – French phrase book and sports dictionary / Inglis – Pisin – Frans tok save na spot diksineri / Anglais – Pidgin – Française dictionnaire phraséologique et sportif. – Rabaul [New Britain, PNG]: Trinity Press.	*Thematic dict. English > French > TP; alphab. dict. English >TP > French.*
24	1969	Mihalic, [Father] Francis (1969): Introduction to New Guinea Pidgin. – Milton [Queensland]: The Jacaranda Press.	*Alph. and thematic dict. Engl. > TP.*
25	1969	Steinbauer, Friedrich (1969): Concise Dictionary of New Guinea Pidgin (Neo-Melanesian) with translations in English and German. – Madang [New Guinea, PNG]: Kristen Pres Inc.	*TP > English > German.*
26	1971	Mihalic, Francis (1971): The Jacaranda Dictionary and Grammar of Melanesian Pidgin. – Milton et al. [Queensland]: Jacaranda Press.	*TP > Engl.; Engl. > TP; thematic word lists English > TP.*
27	1971	Wurm, S[tephen] A. (1971): New Guinea Highlands Pidgin: Course Materials. – Pacific Linguistics, Series D, 3. Canberra: The Australian National University.	*TP > English, ordered according to part of speech.*
28	1973	Balint, Andras (1973): Towards an Encyclopedic Dictionary of Nuginian (Melanesian Pidgin). – Kivung 6, 1-31.	*Articles (letter A) of a proposed monolingual dict. of TP.*
29	1973	Dutton, T[homas] E[dward] (1973): Conversational New Guinea Pidgin. – Pacific Linguistics, Series D, 12. Canberra: The Australian National University, Department of Linguistics, Research School of Pacific Studies.	*Word lists TP > English added to learning units.*
30	1978	Strickert, Frederick (1978): Diksenari Bilong Nupela Testamen. New Testament Dictionary in New Guinea Pidgin. – Madang [New Guinea, PNG]: Kristen Pres.	*Monolingual TP reference work (personal and place names from the Bible).*
31	1985	Dutton, Tom & Dicks Thomas (1985): A new course in Tok Pisin (New Guinea Pidgin). – (Pacific Linguistics, D-67 / Languages for intercultural communication in the Pacific area project of the Australian Academy of the Humanities, 2). Canberra: The Australian National University.	*TP > English; English > TP.*

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Nº	Year	Reference	Comment
32	1985	Murphy, John J. (1985): The Book of Pidgin English. Buk bilong Tok Pisin. – Revised ed. Bathurst: Robert Brown & Assoc.	*TP > English; English > TP.*
33	1986	Schäfer, Albrecht (1986): Pidgin-English für Papua Neuguinea. – (Kauderwelsch, 18.) Bielefeld: Peter Rump Verlag.	*German > TP; TP > German.*
34	1992	Lloyd, J. A. (1992): A Baruya-Tok Pisin-English Dictionary. – Canberra: The Australian National University, Department of Linguistics, Research School of Pacific Studies.	*Baruya > TP > English; TP > Baruya; English > Baruya.*
35	1996	Barhorst, Terry D. & Sylvia O'Dell-Barhorst (1996): Pidgin / English Dictionary as spoken in Port Moresby, Papua New Guinea – http://www.june29.com//HLP/lang/pidgin.html. (last access 01-01-2021).	*Not accessible anymore; English > TP.*
36	1996	Slone, Thomas H. (1996): Tok Nogut. An Introduction to Malediction in Papua New Guinea. – Maledicta: The International Journal of Verbal Aggression 11, 75-104.	*TP > English.*
37	1997	Kocher Schmid, Christin (1997): Terms in Neo-Melanesian for plants and animals. – http://lucy.ukc.ac.uk/rainforest/frp-website/Publications/worksheets/SHEET3/biopidg_1.html.	*Not accessible anymore; TP > English and/or Latin.*
38	1997	Thomas, Dicks R., T. R. Andi Lolo, & Nico Jakarimilena (1997): Trilingual Dictionary Tokpisin English Bahasa Indonesia. – Port Moresby: The Education and Cultural Attache of the Indonesia Embassy, and The Department of Language & Literature, UPNG, [printed by Balai Pustaka, Jakarta].	*TP > English > Indonesian.*
39	2001	Newlin, Andy (2001): Tok Pisin / Pidgin / English Online Dictionary. – http://www.tok-pisin.com/ (last access: 20-03-2022).	*TP > English.*
40	2003 / 2006	Feldpausch, Becky (Hg.) (2003/2006): Almalu kali, Eyo kali, i Walowei luk kal. Namia, Tok Pisin, and English Dictionary. – Revised for website (April 2006). Ukarumpa [New Guinea, PNG]: SIL Press. https://pnglanguages.sil.org/resources/archives/39181 (last access: 19-03-2022). [In print: Ukarumpa, New Guinea, PNG: Summer Institute of Linguistics Press 2003.]	*Namia > TP > English.*
41	2005	Ward, Stephen (2005): Clinical Clerking and Examination in Tok Pisin. A resource for English speaking health care workers in Papua New Guinea. – [Wewak, New Guinea]. http://studylib.net/doc/7538271/clinical-clerking-and-examination-in-tok (last access: 18-03-2022).	*Short word lists English > TP.*
42	2006	Burton, John (2006): Revising the Mihalic project. http://pandora.nla.gov.au/pan/67828/20120213-0001/www.mihalicdictionary.org/index-3.html [last access: 19-03-2022].	*TP > English; collaborative revision of Mihalic (1971).*
43	2006	Garnier, Nicolas (2006): Dictionnaire Français / Tok Pisin. Buk bilong ol nem long Tok Pisin na Tok Franis. – Port Moresby [New Guinea, PNG]: Alliance Française de Port Moresby & the University of Papua New Guinea.	*TP > French; French > TP.*

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Nº	Year	Reference	Comment
45	2006	Lothmann, Timo (2006): God i tok long yumi long Tok Pisin. Eine Betrachtung der Bibelübersetzung in Tok Pisin vor dem Hintergrund der sprachlichen Identität eines Papua-Neuguinea zwischen Tradition und Moderne. – Frankfurt/M.: Lang.	*TP > German.*
46	2007	Pernet, Barbara & Wolfgang Wendt (2007): Tok Pisin bilong Papua Niugini – Das Pidgin von Papua-Neuguinea. Eine Einführung. Sprachkurs in 16 Lektionen. – Neuendettelsau: Mission EineWelt Centrum für Partnerschaft, Entwicklung und Mission der Evangelisch-Lutherischen Kirche in Bayern.	*TP > German; German > TP.*
47	2008	Parker, Philip M. (Hg.) (2008): Webster's Tok Pisin – English Thesaurus Dictionary. – San Diego: ICON Group Intern.	*TP > English; English > TP.*
48	2008	Volker, Craig [general editor], Susan Baing, Brian Deutrom & Russell Jackson (2008): Papua New Guinea Tok Pisin English Dictionary. South Melbourne: Oxford University Press.	*TP > English; English > TP; new ed. in preparation.*
49	2017 ff.	Engelberg, Stefan, Christine Möhrs & Doris Stolberg (2017 ff.): Wortschatz deutschen Ursprungs im Tok Pisin. Version 1. – In: Meyer, Peter & Stefan Engelberg (2012 ff.): Lehnwortportal Deutsch. Mannheim: IDS. http://lwp.ids-mannheim.de/doc/tokpisin/start (last access: 22-03-2022).	*TP > German; dynamically published.*
50	2017 ff.	Anonym (2017 ff.): Tok Pisin English Dictionary. Tok Pisin (New Guinea Pidgin) English Bilingual Dictionary & Encyclopedia of Papua New Guinea. – https://www.tokpisin.info/ (last access: 21-03-2022).	*TP > English; English > TP; mainly an unlicensed version of Volker et al. 2008.*

Table 1: Chronologically ordered list of the dictionaries in the Tok Pisin Dictionary Collection of the Leibniz Institute for the German Language [*TP = Tok Pisin*]

Some other early dictionaries on Melanesian pidgins like Churchill (1911) and Pionnier (1913) are not listed here as they do not refer to the New Guinean pidgin variant.

3. Colonial phases and colonial actors in the lexicographic history of Tok Pisin

Lexicography prior to Papua New Guinea's independence is part of colonial history. Lexicographic strategies, dictionary structures, and publication patterns reflect the interest (and disinterest) of various colonial actors in the Tok Pisin language area. The particular case of Tok Pisin also demonstrates the role dictionaries play in a colonial society.[3]

Three main groups of colonial actors are identified as promoters and practicioners of dictionary making with respect to Tok Pisin: European scientists (section 3.1), Catholic missionaries (section 3.2), and individuals associated with the Australian and American militaries (section 3.3). Figure 1 shows how the dictionaries produced by these three groups are distributed over the time between 1900 and 1975.

[3] Contributions on parts of the lexicographic history of Tok Pisin have been made by Laycock (1977), Mühlhäusler (1985a, 1985b), and Engelberg/Stolberg (2017).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

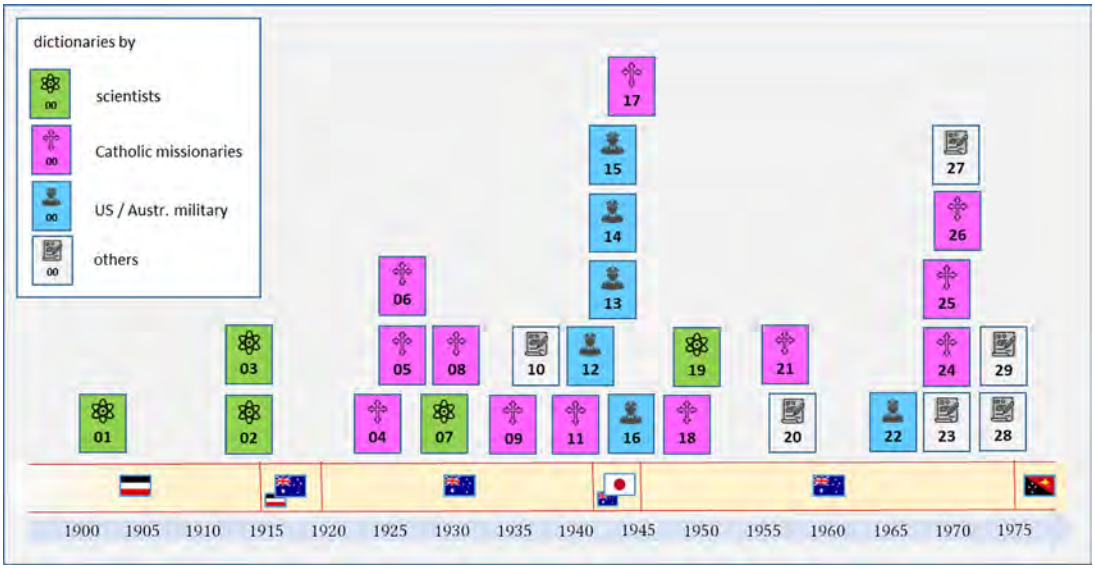


Fig. 1: Temporal distribution of dictionaries of Tok Pisin according to the main groups of lexicographers (numbers refer to table 1)

3.1 European scientists as lexicographers

Between the 1880s and 1920s, an enormous number of scientific books and articles were published in German based on research in and about German New Guinea in the fields of anthropology, linguistics, medicine, agriculture, geography, geology, and so on. However, those who conducted fieldwork were confronted with a multitude of some 800 languages spoken in what is now Papua New Guinea. When traveling, chains of interpreters were often required to ensure communication. As Tok Pisin spread within the Bismarck Archipelago in the late 19th century and further into "Kaiser-Wilhelmsland" (northeastern New Guinea) and the northern Solomon Islands, explorers and traveling scientists began to use Tok Pisin as a lingua franca. Some of them, such as Otto Dempwolff (cf. Fig. 2), Richard Thurnwald, and Beatrice Blackwood (cf. Fig. 3), began to compile small, mostly handwritten dictionaries of Tok Pisin, primarily for their own research purposes.

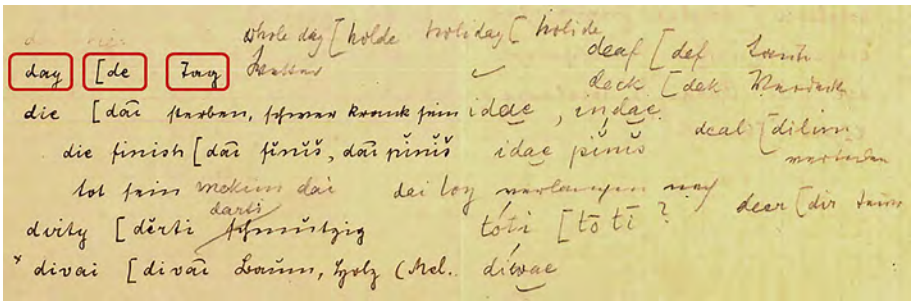


Fig. 2: Dempwolff`s ([ca.] 1902/1913) dictionary "Pidgin-Englisch von Deutsch-Neuguinea" with a microstructure comprising of a Tok Pisin lemma in etymologizing English lexicography (e. g. *day*), a phonetic representation of the lemma sign ([*de*), and a German equivalent (*Tag*)

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

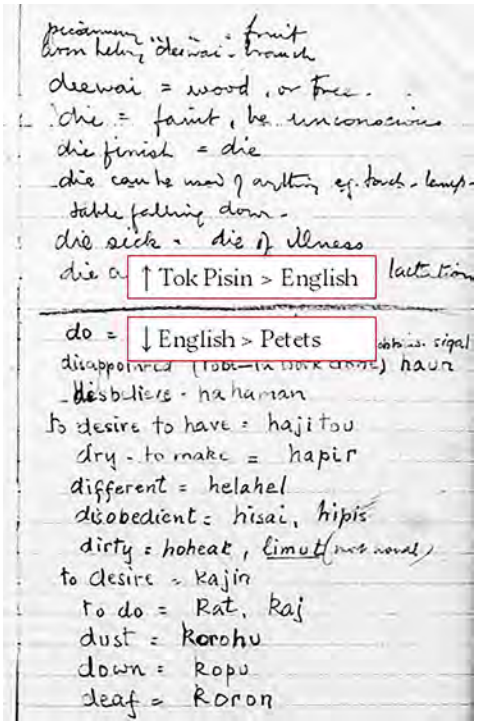


Fig. 3: Blackwoods's (1929) "Pidgin & English-Petets Vocabulary" in which the dictionary English to Petets (spoken in the Northern Solomon Islands) is preceded with a small Tok Pisin to English word list for each letter, indicating that Tok Pisin was used for communication in fieldwork

3.2 Lexicography in Catholic missions

The Catholic missions, dominated by German missionaries until World War II, switched from local languages (in part) to German and, in the late 1920s, to Tok Pisin as the mission language.

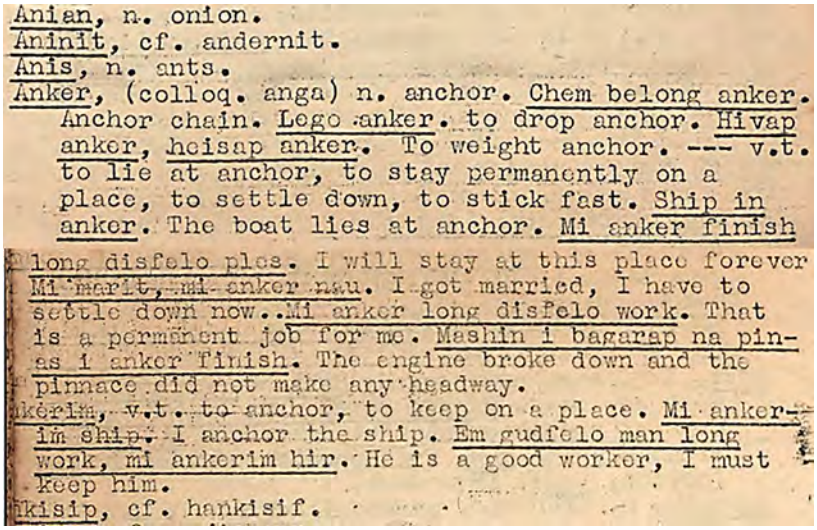


Fig. 4: Dictionary articles for Tok Pisin *anker/anka* ('Anker') and *ankerim* ('ankern') in Schebesta/Meiser's (1945) "Dictionary of "Bisinis English" (Pidgin-English)", giving an impression of the extensive microstructure containing information on word variants, collocations, multiword expressions, example sentences, and metaphoric uses

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

This process was accompanied by extensive lexicographic work. However, the resulting dictionaries were not published for several decades, but only used internally by the mission. It was not until 1957 – after eight unpublished mimeographed dictionaries, some of them quite substantial – that the first dictionary by a Catholic missionary went to press (Mihalic 1957). Given the extensive activities of the missions (mission work, education, plantation operations, shipping, crafts, etc.), these dictionaries were comprehensive and they had quite a high lexicographical quality (cf. Fig. 4).

In contrast to the Catholic missions, the Protestant missions hardly dealt with Tok Pisin lexicographically. Card index boxes containing index cards for several thousand Tok Pisin lemmas (Jäschke [ca.] 1955–1960), which have recently surfaced from the archives of the Neuendettelsau Mission, represent the only lexicographic output of the Protestant missionaries in our dictionary collection.[4] This is probably mainly due to the different attitude of the Protestant missions towards the indigenous languages, which goes back to Luther's dictum that the Bible should be rendered in the vernacular languages. Thus, the Protestant missions relied more heavily on the local languages in their missionary work.[5]

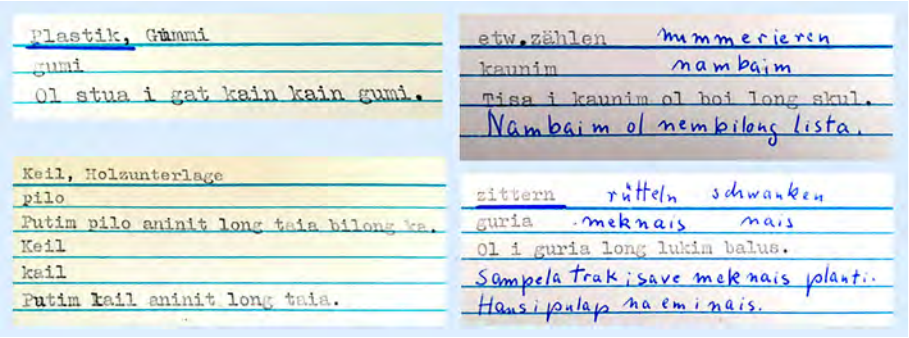


Fig. 5: Index cards from Jäschke's ([ca.] 1955-1960) lexicographic project where handwritten addenda illustrate the state of revision

3.3 Lexicography in times of war

Neither under German nor under Australian rule (after WW I) did Tok Pisin find official support. In German New Guinea, Europeans regarded Tok Pisin as a corrupted form of English, and the use of Tok Pisin was officially disapproved by the German administration. Despite this attitude, Tok Pisin was increasingly used by the local population, by settlers and traders, and even by the German administration itself. Tok Pisin thus played a crucial role in the German colony. However, attempts at a lexicographic description were neither undertaken nor supported by the German government.

Under Australian rule, attitudes did not change significantly. Australians expected Tok Pisin to be replaced by English in the long run. With the onset of WW II, however, cultural-political considerations took a back seat when social control, propaganda, and efficient communication between the military and civilians became necessary. Native speakers of English

[4] The dictionary is currently subject of a student thesis written by Melanie Drothler.

[5] For the history of the missions in New Guinea and the question of the mission language, cf., e. g., Eggert (1997) and Steffen (2001).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

associated with the Australian or U.S. military therefore produced mostly simple dictionaries mainly for military personnel (cf. Fig. 6).

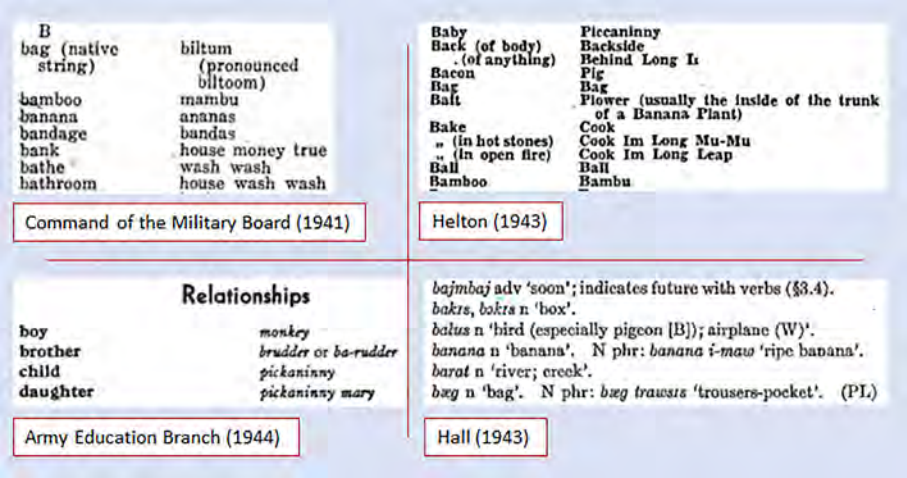


Fig. 6: Excerpts from dictionaries connected to the Australian military (Command of the Military Board 1941; Helton 1943) and to the US army (Hall 1943; Army Education Branch 1944)

After World War II, Tok Pisin was used as the language of instruction in some schools. Therefore, in the mid-1950s, the Department of Education pushed for the development of a standard orthography for Tok Pisin. The orthography approved by the Department of Education was then used in Mihalic's (1957) dictionary (Wurm 1985, pp. 170 f.).

4. Post-independence lexicography of Tok Pisin

In summary, prior to independence, Tok Pisin lexicography was practiced by expatriates in order to explore New Guinea scientifically, missionize the population, and wage a war whose causes and goals were alien to the indigenous population. However, in the wake of independence, as New Guineans more and more gained sovereignty in matters of language policy, Tok Pisin expanded its functional domains as spoken language (public life, parliament, radio) and also became more visible in its written form. The early editions of the "Wantok Niuspepa" from 1970 onwards still provide a vivid picture of this period and the role that written Tok Pisin played in this process (Wantok 2014 ff.). Rather inadvertently, missionary lexicography supported this process of regaining linguistic self-determination.

Catholic missionaries had continued to produce dictionaries, which were printed between the late 1950s and the early 1970s. In 1971, Mihalic's "Jacaranda Dictionary" was published, the scope and quality of which brought general lexicography to a standstill for a long time. After Papua New Guinea's independence in 1975, Tok Pisin lexicography was limited to dictionaries and word lists with specific functions, such as the travel dictionary by Schäfer (1985) or the etymological dictionary by Engelberg/Möhrs/Stolberg (2017 ff.), and to dictionaries and word lists of specialized language, e. g., in the fields of religion (Strickert 1978; Lothmann 2006), medicine (Ward 2005), and botany (Kocher Schmid 1997). New editions of older dictionaries (Murphy 1985), word lists in language courses (Dutton 1985; Pernet/Wendt 2007), and bilingual and trilingual dictionaries with languages from the region (Thomas/Lolo/Jakarimilena 1997; Lloyd 1992; Feldpausch 2003/2006) complete the picture. It is only in the 21st century that new general dictionaries of Tok Pisin are being published

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

(Garnier 2006, and especially Volker et al. 2008). Since the 1990s, Internet lexicography has also taken hold, producing new dictionaries with mostly simple microstructure (Barhorst/O'Dell-Barhorst 1996; Newlin 2001), unlicensed digital versions of printed dictionaries (e.g., Anonymous 2017 ff., based on Volker et al. 2008), and an excerpted revision of Mihalic's work (Burton 2006).

A comprehensive history of Tok Pisin lexicography in the post-independence period has yet to be written. It would have to take into account the interaction of lexicography with educational and language policies, the role of the missions, the complex dual structure of traditional power relations and modern state institutions, and, of course, the functions of English, Tok Pisin, and the many „tokples" (indigenous languages) in Papua New Guinea. And the history of recent Tok Pisin lexicography will have to explain why post-independence lexicography is still dominated by expatriates, and why, as one reviewer asked, „the lexicographic space in this supposedly independent country is still not filled by indigenous voices".

References

Baker, P. (1993): Australian influence on Melanesian Pidgin English. In: Te Reo 36, pp. 3–67.

Churchill, W. (1911): Beach-la-mar. The jargon or trade speech of the western pacific. Washington DC.

Eggert, J. (1993): Aspekte der Sprachenfrage in der Arbeit deutscher Missionen in Neuguinea. In: Verbum SVD 34, pp. 283–301.

Engelberg, S. (2014): Die deutsche Sprache und der Kolonialismus. Zur Rolle von Sprachideologemen und Spracheinstellungen in sprachenpolitischen Argumentationen. In: Kämper, H./Haslinger, P./Raithel, T. (eds.): Demokratiegeschichte als Zäsurgeschichte. Berlin/Boston, pp. 307–332.

Engelberg, S./Stolberg, D. (2017): The influence of German on the lexicon of tok pisin. In: Language and linguistics in Melanesia, special issue 2017 [Maitz, P./Volker, C. A. (eds.): Language contact in the German colonies: Papua New Guinea and beyond], pp. 27–64. https://langlxmelanesia.com/special-issues (last access: 20-03-2022).

Laycock, D. C. (1977): A history of lexicography in the New Guinea area. In: Wurm, S. A. (ed.): New Guinea area languages and language study. Canberra, pp. 169–192.

Mühlhäusler, P. (1978): Samoan plantation pidgin English and the origin of New Guinea pidgin. In: Papers in Pidgin and Creole Linguistics 1, pp. 67–119.

Mühlhäusler, P. (1979): Growth and structure of the lexicon of New Guinea Pidgin. Canberra.

Mühlhäusler, P. (1985a): History of the study of Tok Pisin. In: Wurm, S. A./Mühlhäusler, P. (eds.): Handbook of Tok Pisin (New Guinea Pidgin). Canberra, pp. 15–33.

Mühlhäusler, P. (1985b): Kritische Bemerkungen zu Wörterbüchern des Tok Pisin und anderer Kreolsprachen. In: Boretzky, N./Enninger, W./Stolz, T. (eds.): Akten des 1. Essener Kolloquiums über „Kreolsprachen und Sprachkontakte" vom 26.1.1985 an der Universität Essen. Bochum, pp. 71–85.

Pionnier, J.-N. (1913): Pigeon English ou Bichelamar. In: Revue de linguistique et de philologie comparé 46, pp. 109–117, pp. 184–198.

Steffen, P. (2001): Die katholischen Missionen in Deutsch-Neuguinea. In: Hiery, H. J. (ed.): Die deutsche Südsee 1884–1914. Ein Handbuch. Paderborn/München/Wien/Zürich, pp. 343–383.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Wantok (2014 ff.): Wantok niuspepa – archives. Word Publishing Company LTD. https://wantokniuspepa.com/index.php/archives/wantok-niuspepa (last access: 20-03-2022).

Wurm, S. A. (1985): Writing systems and the orthography of Tok Pisin. In: Wurm S. A./ Mühlhäusler, P. (eds.): Handbook of Tok Pisin (New Guinea Pidgin). Canberra, pp. 167–176.

Contact information

Stefan Engelberg
Leibniz-Institut für Deutsche Sprache
engelberg@ids-mannheim.de

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/ Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Laura Giacomini/Paolo DiMuccio-Failla/
Patrizio De Martin Pinter

THE REPRESENTATION OF CULTURE-SPECIFIC LEXICAL ITEMS IN MONOLINGUAL LEARNER'S LEXICOGRAPHY

The case of the electronic Phrase-Based Active Dictionaries

Abstract This paper focuses on the treatment of culture-bound lexical items in a novel type of online learner's dictionary model, the Phrase-Based Active Dictionary (PAD). A PAD has a strong phraseological orientation: each meaning of a word is exclusively defined in a typical phraseological context. After introducing the relevant theory of *realia* in translation studies, we develop a broader notion of culture-specific lexical items which is more apt to serve the purposes of learner's lexicography and thus to satisfy the needs of a larger and often undefined target group. We discuss the treatment of such words and expressions in common English learner's dictionaries and then present various excerpts from PAD entries in English, German, and Italian which display different strategies for coping with cultural contents in the lexicon. Our aim is to demonstrate that the phraseological approach at the core of the PAD model turns out to be extremely important to convey cultural knowledge in a suitable way for users to fully grasp cultural implications in language.

Keywords Learner's lexicography; phraseology; culture-specific items; realia; multimedia

1. Introduction

This contribution concentrates on the treatment of words and expressions indicating culture-specific items in the model for online Phrase-based Active Dictionaries (PADs), which is currently being implemented in the context of the PhraseBase project. PhraseBase is a new project in learner's lexicography carried out at the Universities of Hildesheim and Heidelberg. At the core of the project are a cognitive approach to the study of language, a strong phraseological orientation towards lexical analysis and representation, as well as a corpus-based method to data acquisition and preparation. The theoretical background and the lexicographic process have been described in DiMuccio-Failla/Giacomini (2017a, 2017b), Giacomini/DiMuccio-Failla (2019), Giacomini/DiMuccio-Failla/Lanzi (2020), as well as in forthcoming publications. The current state of the PADs is a set of distinct monolingual resources (English, Italian, German) in which several sample entries have been compiled, especially for verbs. Each PAD entry has a deep hierarchical structure in which collocations are systematically employed at each level as disambiguating elements and possibly as components of normal patterns of usage (Sinclair 2004; Hanks 2013).

The online PADs are mainly addressed at non-native speakers of a language and cover data suitable for the CEFR levels B1-C2 that can be selectively presented according to the profile of the individual user. The distinctive cognitive linguistic character of the project is reflected, among others, by the access structures and the microstructure of the dictionaries, including the applied sorting criteria, and the treatment of polysemy, with the identification of progressive extensions of core prototypical meanings. Alongside general linguistic-phra-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

seological aspects, the cultural dimension is essential. It manifests itself in the lexicon in various ways, for example in the denotative or connotative meaning of words, in the use of familiar alternative terms, in the pragmatic nuances of language.

Words and expressions characterised by cultural specificity pose a challenge for non-native speakers, e.g. foreign language learners or translators, both during text reception and text production tasks. Their role in learner's lexicography has already been discussed in a number of publications concerned with different languages and different subtypes of dictionaries (cf., among others, Zhang/Mi 2020; Tomaszczyk 2017). In the context of translation studies, culture-bound words and expressions, often referred to as *realia*, have been extensively explored and are mostly analysed from the point of view of distinct language pairs (cf. House 2004; Markstein 1998; Schreiber 2007). We will draw on these studies but attempt to devise a description model that is suitable for monolingual learner's lexicography, in which a range of possible target languages and cultures needs to be taken into account. The goal of this contribution is to show how the analysis and presentation of words denoting culture-specific items can benefit from PhraseBase's focus on the phrasal nature of language in terms of information completeness and potential information delivery efficiency.

After discussing the way in which existing learner's dictionaries treat this kind of data and illustrating different types of cultural specificity in the general language lexicon, the following aspects will be considered in greater detail: principles for selecting culture-bound vocabulary, its integration in the PAD microstructure and the definition of adequate microstructural data types, including multimedia options. This will be illustrated by way of examples in Italian, German, and English. Our final goal is to demonstrate that the phraseological approach at the core of the PAD model is crucial to convey cultural knowledge in a suitable way for dictionary users to fully grasp cultural implications in language.

2. Theory on culture-specific lexical items: from translation studies to learner's lexicography

In this contribution, *culture-specific* is broadly intended as the property of lexical items related to non-universal concepts (and entities) and displaying peculiar socio-cultural, historical, geographical etc. meaning traits. In principle, any type of dictionary may contain words and expressions related to culture-specific entities. In a general monolingual dictionary aimed at native speakers, such words and expressions usually have the same status as any other element of the lexicon. An exception is when cultural specificity is linked, for example, to certain characteristics of a restricted geographical area of a country and to certain diatopic varieties of a language (e.g. *carsico* in Italian and *Apfelsine* in German). In that case, short encyclopaedic additions to definitions as well as pragmatic indications are useful to support text comprehension.

Cultural specificity can only be grasped when different cultural and linguistic realities are compared. This is even more apparent when dealing with bilingual dictionaries and learner's dictionaries. Here, words and expressions related to culture-specific entities should, at least in theory, be considered as a part of the lexicon that requires special lexicographic treatment, precisely because it is designed for a potential non-native user with a (partly) different cultural background.

It is not surprising that many reflections on the nature of such elements of the lexicon and the problems they pose for a non-native speaker have originated in the field of translation

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

studies. In this context, a common term used to indicate culture-specific lexical items is *realia*. This term was first used in this meaning by translation scholars from Eastern Europe and the Soviet Union in the 1950s and 60s, such as Sobolev in 1952, and Vlachov/Florin in 1969 and then in 1980. According to Sobolev, *realia* are words and phrases specifically related to the everyday life of a nation, which do not have any equivalents in the everyday life – and therefore – in the languages of other countries (Sobolev 1952, p. 281). A decade later, the Bulgarian translation scholars Vlachov and Florin expanded Sobolev's theory in two monographs published, respectively, in 1969 and 1980, and defined *realia* as

> [...] words (and phrases), denoting elements that are specific to the life (everyday life, culture, social and historical progress) of a people and foreign to other peoples; since they carry a national and / or historical connotation, they do not have, as a rule, precise correspondences (equivalents) in other languages, and therefore cannot be translated «following the common rules» of translation, but require a specific approach.[1] (Vlachov/Florin 1980, 47; translation by P.D.M.P.)

Their work *Neperevodimoe v perevode* ("The Untranslatable in Translation") published in 1980 represents one of the few scientific works which have investigated the concept of *realia* in such an extensive way. Moreover, to the best of our knowledge Vlachov and Florin are the only authors who have attempted to propose a categorization of *realia* (cf. Section 3). Being aware of the impossibility of establishing well-defined borders between *realia* and other categories, such as proper nouns, terms, and appellatives, they also provided interesting differentiation criteria with several examples.

In the wake of Vlachov and Florin, scholars belonging to the Leipzig school of translation such as Otto Kade, Albert Neubert and Wladimir Kutz carried on the research about *realia* focusing on viable translation strategies, especially in the language pair Russian-German (cf. Kutz 1977). In the last decades of the 20th century, the interest in *realia* grew also in Western Europe and North America, as other researchers (cf. Newmark 1981; Williams 1990; Kujamäki 2004) began to focus their studies on how to deal with *realia* in the translation process within the scope of a specific language pair.

The definitions mentioned above, as well as those formulated by Bödeker/Freese (1987) and Koller (1979), allow for a very precise approach to the translation of realia but fail to provide objective criteria for their identification outside the translation process, i. e. without the comparison between a source language and a target language.

Markstein attempts to provide an objective interpretation of *realia* and defines them as elements of the everyday life, history, culture, politics, etc. of a given people, country, geographical place, which do not have any correspondences in other peoples, other countries, other geographical places (Markstein 1998, p. 288). She also applies a decisive criterion and states that *realia* are identity carriers of a national/ethnic entity, a national/ethnic culture and are associated to a country, a region, a continent (ibid). An important aspect of her definition is based on a specific distinction: *realia* are objects, phenomena, etc. which have no correspondence in other *cultures*, and not just in other *languages*.

[1] Original text (Russian): "[...] слова (и словосочетания), называющие объекты, характерные для жизни (быта, культуры, социального и исторического развития) одного народа и чуждые другому; будучи носителями национального и/или исторического колорита, они, как правило, не имеют точных соответствий (эквивалентов) в других языках, а, следовательно, не поддаются переводу «на общих основаниях», требуя особого подхода."

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

The problem of lexical equivalence is, of course, crucial in the context of translation studies. The definition provided by Markstein, however, opens a broader perspective on the role of cultural differences which is more apt to serve the purposes of a learner's dictionary. This aspect will be discussed in the next section.

3. Types of culture-specific lexical items

The categorization of *realia* proposed by Vlachov/Florin (1980, pp. 47–79) stretches over multiple levels of categorization and implies that a *reale* can simultaneously belong to more than just one of them. The first level of categorization concerns the denotatum: the authors subdivide *realia* into geographical (e. g. pampas, fjord, yeti, sequoia, etc.), ethnographic (e. g. sauna, kimono, sombrero, balalaika, etc.), and socio-political realia (e. g. Tory, tsar, sheikh, pharaoh, etc.). The second level of categorization takes into account the language(s): *realia* can be analysed either as elements of a single language, and therefore, be subdivided in domestic and foreign *realia*, or as contrasting elements between two languages. In this case, *realia* can be internal or external to the translation language pair. Moreover, according to a third level of categorization, *realia* can be subdivided diachronically into historical and contemporary *realia*. A comparable categorization has been made by Newmark (1988, p. 95). Newmark identifies the following cultural categories: 1) ecology; 2) material culture; 3) social culture; 4) organization, customs, ideas; 5) gestures and habits. A last differentiation criterion pointed out by Vlachov and Florin and pertaining to the present study involves the lexicographic occurrence of a *reale*: lexicographers should always take into consideration the frequency of a foreign *reale*, whether it has enjoyed an ephemeral success in a given language or whether it still fulfils the criteria for being defined as such.

As already mentioned, we draw our inspiration from studies on cultural categories and *realia* in translation, however we do not specifically concentrate on words and expressions for which no target notion (i. e. a referent in a target culture) and no equivalents exists. In fact, the problem of equivalence itself will not be dealt with in this study. From the standpoint of learner's lexicography, we need to set up a more flexible approach to the definition of cultural items and their treatment in a dictionary for learners. We expand the translation-oriented perspective by taking into account, among others, the approach by Markstein (1998) and the concept of *cultureme* in Pamies (2017) to include more phenomena than mere lexical gaps and pay close attention to the linguistic and encyclopaedic needs of a much broader and partially undefined public. The broader picture is given by the interplay between concepts and their lexicalisations in different cultures, for which a range of cases can be identified (Table 1).

1. Referent is available in a certain target culture:	**2. Lexicalisation of the related concept in the target language:**	**3. Difference between the word in the source and in the target language in terms of cultural content:**
No	–	–
Yes	Not (yet) lexicalised	–
	Lexicalised – sometimes through a foreign word	– Culturally different – Culturally similar

Table 1: Concepts and their lexicalisations in different cultures

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

It should be the goal of a learner's dictionary that addresses users of many different languages and cultures to provide specific cultural information related to the lexicon, especially when it can be assumed that some lexicalised concepts may be experienced differently in different cultural environments. This is often the case of foreign words in a language: is a *pub* in Italy the same as a *pub* in the UK, or is the *Bolognese-Sauce* in Germany the same as the *ragù alla bolognese* in Italy? Of course, these borrowings need to be treated like neologisms which may slowly adapt their meaning to the cultural adaptations of the referent. This also should be reflected by an up-to-date lexicographic resource. Cultural differences, finally, can depend on encyclopaedic, i. e. non-linguistic, but also be paired to differences in connotative meaning (think, for example, of the connotation of words indicating the colour red in different languages).

4. The treatment of culture-specific lexical items in learner's dictionaries

As previously mentioned, the treatment of culture-specific words in lexicography is particularly important for a non-native speaker. In a bilingual dictionary, such treatment can effectively focus on the language pair in question, depending also on the directionality of the dictionary (cf. Svénsen 2009). A monolingual learner's dictionary typically does not target a single group of native speakers. The presentation of culturally specific vocabulary elements should be therefore calibrated with a potential audience in mind, possibly including different language families. The number of lexical items of a language which can be seen as culture-specific is determined by the distance between the socio-cultural context of that language and that of other languages. This topic has been dealt with in studies on equivalence in translation (cf., among others, Koller 1979).

In this section we offer a brief overview of how four online monolingual English learner's dictionaries[2] treat lexical items which are probably unfamiliar to most of their dictionary users because of their cultural specificity. For each of the three languages, three common words and expressions have been selected and analysed from the perspective of their 'cultural', encyclopaedic coverage. The headwords are words indicating concrete or abstract concepts, including actions and events. In this test, we aimed at choosing words and expressions which are not being commonly used as foreign words in other languages (i. e. not such cases as *pub* or *paella* in Italian). *Fish and chips*, *peer*, *hustle and bustle*, and *to facepalm* have been selected to show different ways in which cultural aspects may be rooted in language. For analysing dictionary data, we applied the following criteria:

- adequacy of semantic information: do definitions provide all necessary information for the user to grasp the cultural implications behind the meaning(s) of the lexical item?
- adequacy of pragmatic information: do lexicographic items such as pragmatic labels and usage examples enable the user to fully understand the (cultural) context in which a lexical item is used?

We intentionally concentrate on the quality of data needed for text reception alone, since this is the most important step for the user to first deal with an unfamiliar part of the lexicon. Grammatical aspects are not taken into account, as they do not typically have a specific

[2] Longman Dictionary of Contemporary English (LDOCE), Oxford Advanced Learner's Dictionary (OALD), Cambridge Advanced Learner's Dictionary (CALD), and Collins English Dictionary (COLLINS).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

impact on the treatment of this kind of lexical units if compared with others. Table 2 lists the definitions provided in the selected dictionaries:

fish and chips:	[LDOCE] – a meal consisting of fish covered with batter (=a mixture of flour and milk) and cooked in oil, served with long thin pieces of potato also cooked in oil [OALD] – a dish of fish that has been fried in batter served with chips, and usually bought in the place where it has been cooked and eaten at home, etc., **especially in the UK** [COLLINS] – Fish and chips are fish fillets coated with batter and deep-fried, eaten with French fries. [CALD] - fish covered with batter (= a mixture of flour, eggs, and milk) and then fried and served with pieces of fried potato
peer:	[LDOCE] – a member of the **British** nobility ⟶ House of Lords, peerage [OALD] – **(in the UK)** a member of the nobility or the House of Lords [COLLINS] – **In Britain**, a peer is a member of the nobility who has or had the right to vote in the House of Lords [CALD] – **in the UK**, a person who has a title and a high social position
hustle and bustle:	[LDOCE] [OALD] [COLLINS] – (no entry, there are only examples) [CALD] – busy movement and noise, especially where there are a lot of people (separate treatment in the entry of hustle)
to face-palm/ facepalm:	[LDOCE] – (informal) to put the palm of your hand on your face, or put your face down on your hand, when you are embarrassed, disappointed, shocked at someone's stupidity etc [OALD] – (informal) the action of covering your face with your hand, usually because you are shocked, embarrassed, annoyed, etc. [COLLINS] – to bring the palm of one's hand to one's face as an expression of dismay [CALD] – (informal) to cover your face with your hand because you are embarrassed, annoyed, or disappointed about something

Table 2: Definitions of culture-specific lexical items in four major learner's dictionaries of English

With the exception of some indications which we highlighted by using bold characters, definitions are neutral from the point of view of cultural specificity and the degree of possible semantic-encyclopaedic coverage varies from one dictionary to the other.

Fish and chips indicates a concrete entity, a dish or even a meal, which is generally described by external sources such as encyclopaedic and news sources as being typically British. Its typicality is also reflected by the fact it is a multiword expression. Although the meaning of the expression is transparent and fully compositional, and the definitions provided by the dictionaries are exhaustive from the perspective of the purely denotative meaning, subtle connotative information related to the context in which this kind of dish is sold and eaten is missing. The same observations apply to *peer*, a word which describes a specific role in British society and politics.

In general, no images are provided for the selected words, even though this would be a useful device for supporting text reception. The only exception is CALD's image for *fish and*

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/ Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

chips. CALD also mentions a large number of related words and phrases in a dictionary section called SMART vocabulary, in which further culture-specific lexical items indicating dishes are listed, for instance *baked beans*, *arancini*, *French toast*, *pad thai*, and *spaghetti bolognese*.

The specificity of the idiom *hustle and bustle* seems to have been fully underestimated and CALD is the only dictionary that provides a definition. The verb *to facepalm*, finally, indicates a gesture and, as such, can turn to be highly culture-specific depending on the recipient. At first sight, the definitions in the four dictionaries are similar, but the description of the action involved in facepalming oneself can be pretty different, ranging from generic "putting the palm of the hand on one's face" to "covering one's face with the hand". In order to specify this kind of action, some images would have been useful. There are, in fact, different ways of performing this action: with one or two hands as well as putting your hands in different and yet typical positions on your face, often depending on the emotion you are experiencing (embarrassment or disappointment, for example). Whenever facial expressions or, in general, non-verbal communication are involved, the use of non-verbal descriptors such as images and videos should be considered as an essential complement to definitions.

Pragmatic information is primarily delivered by means of usage examples. The quantity of examples greatly varies from dictionary to dictionary. LDOCE offers the largest number of examples, which seems to be a most adequate solution for illustrating the use of the selected words and expressions in their typical contexts. Especially corpus examples, however, which are provided in addition to a couple of introductory examples, do not really help the dictionary user, since they are either too generic, or too specific, or their context is unclear. Pragmatic labels have not been used with the exception of the *informal* register label for the verb *to facepalm*. The entry for *peer* in LDOCE is related via a link to the topic 'Government'. The webpage dedicated to this topic, however, mostly concentrates on the word *government* itself and makes no mention of the word *peer*.

A comparable treatment of culture-specific lexical items can be found in learner's dictionaries of German (the words *Fachwerkhaus*, *Kabelsalat*, *Schadenfreude*, and *gemütlich* were searched for in Langenscheidt Deutsch als Fremdsprache and PONS Deutsch als Fremdsprache), French (the words *yaourt* and *flâneur* were searched for in the Dictionnaire Le Robert Micro and Le Robert & CLE International), and Spanish (the words *madrugada* and *sobremesa* were searched for in VOX Diccionario de español para extranjeros and ELE Diccionario de español para extranjeros).

Although no large representative study has been carried out in the four languages, these first results seem to confirm that no special focus is put on cultural aspects of the lexicon and their implications in lexicography, be it at the level of semantic or at the level of pragmatic description. Situational knowledge is also rarely conveyed, and non-verbal descriptors such as images are usually missing.

5. Culture-specific lexical item in a Phrase-Based Active Dictionary (PAD)

From the point of view of its lexicographic functions, a Phrase-Based Active Dictionary is primarily intended as a dictionary for text production. However, a prerequisite for enabling

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

the active usage of language is adequate text reception, which is particularly important whenever specific cultural aspects are mirrored in word meanings.

How do we intend to treat cultural specificity in our model for a PAD? The main microstructural feature of the dictionary model is that words are not described in isolation but within typical phraseological patterns, which are seen as the true lexical units of a language. Each phraseological pattern is a syntactic-semantic unit matching a specific word sense, mostly in a collocational way. For instance, we do not just define all possible senses of the verb *to agree*, but we define *to agree with a certain opinion* as the phraseological pattern which uniquely identifies the meaning "to think that a certain opinion is right".

As pointed out in section 3, we widen the typical translation-oriented perspective on culture-bound elements of the lexicon and include a more comprehensive range of phenomena. However, our primary focus is on the base vocabulary of a language, which often has large cross-cultural, if not universal validity. The choice of entries which need to be treated as culture-specific is made by concentrating on lexical elements that on average may be perceived as 'foreign', e. g. their referent is missing or the related concept has not (yet) been lexicalised (cf. Table 1). This is a manual procedure aided by the consultation of existing general dictionaries and the analysis of frequency data in corpora[3].

For some of the selected words and expression, encyclopaedic information is required to provide the dictionary user with an exhaustive cultural picture. Encyclopaedic information can be part of a definition, enclosed in collocations and usage examples, or presented in a dedicated microstructural section. Physical objects shall be complemented by one or more prototypical images. A few entry excerpts will now be presented and discussed from the perspective of their culture-related content:

1) *Fachwerkhaus* in the German PAD:

The excerpt from the entry for *Fachwerkhaus*[4] displays several peculiar features. First of all, corpus-based analysis reveals that the word is mostly used in its plural form and in the context of the description of (usually the centre of) a settlement such as a village, a town, or possibly the old town district of a city. This is reflected by the formulation of the Lexical Unit (in German: Lexikalische Einheit, LE) *(normalerweise alte/historische) Fachwerkhäuser (im Zentrum) einer bestimmten Siedlung* ("(usually old/historical) half-timbered houses (in the centre of) a certain settlement"), in which the semantic type *Siedlung* encompasses different possible types of settlements. The lexical unit functions as a phraseological pattern which is typical for the word *Fachwerkhaus*. It is followed by a definition, usage examples and collocations. A special microstructural section dedicated to culture-specific aspects is called Culture (in German: 'Kultur) and contains encyclopaedic, i.a. historical, architectural and socio-cultural information related to the concept of *Fachwerkhaus*. Two images of pro-

[3] In this study, the English Web 2020, German Web 2018, and Italian Web 2016 corpora have been analysed by using different Sketch Engine tools.

[4] The German word Fachwerkhaus, usually translated in English as half-timbered house, indicates a house built by half-timbering. "Half-timber work was common in China and, in a refined form, in Japan and was used for domestic architecture throughout northern continental Europe, especially Germany and France, until the 17th century." (https://www.britannica.com/technology/half-timber-work)

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

totypical half-timbered houses in the German-speaking area are also included. The Culture section and the images are intended to convey culture-specific information which cannot be given elsewhere. The former, in particular, can include both an entry-specific encyclopaedic description and links to external sources such as Wikipedia, BabelNet, and also popular scientific resources.

5

LE: (NORM. **alte/ historische**) **Fachwerkhäuser** **(im Zentrum) einer best. Siedlung**

DEFINITION: (NORM. alte/ historische) Häuser mit Struktur aus Holzbalken (im Zentrum) einer best. Siedlung

BEISPIELE: 1. *Mehr als 1300 Fachwerkhäuser bilden den historischen Kern der Stadt.* 2. *Bis 1937 war die Schule in einem 1850 erbauten Fachwerkhaus untergebracht.*

KOLLOKATIONEN: *restaurierte/ denkmalgeschützte Fachwerkhäuser; Fachwerkhäuser in der Altstadt/ im Ortskern*

■ KULTUR:

Ein Fachwerkhaus besteht aus einem Art Skelett aus verstrebten Holzbalken. Dieses Stützskelett trägt das gesamte **Fachwerkhaus**. Die jeweiligen Zwischenräume, auch Gefach genannt, sind mit einem lehmbeworfenen Holzgeflecht ausgefüllt oder mit Back-oder Bruchsteinen ausgemauert. Sie können aber auch mit Lehmbausteinen verbaut oder verputzt sein. Verwendete Holzarten sind Eiche (am meisten eingesetzt) oder gelegentlich auch Tanne. (https://www.hausbau-portal.net)

2) *aceto balsamico* in the Italian PAD:

It is important to note that encyclopaedic information can sometimes be paired with information on norms and standards regulating a given domain. This is, for instance, the case of specifications such as DOP, DOC, DOCG, and IGP[6] for Italian food and drinks. In the entry for *aceto balsamico* we include the following data:

5 https://commons.wikimedia.org/wiki/File:Fachwerkh%C3%A4user_%22Am_Johanniskloster%22-20151029-IMG_0516.jpg, https://commons.wikimedia.org/wiki/File:Fachwerkh%C3%A4user_in_Wetter_(Hessen).jpg.

6 DOP: Denominazione di Origine Protetta (Protected designation of origin), DOC: Denominazione di Origine Controllata (Controlled designation of origin), DOCG: Denominazione di origine controllata e garantita (Controlled and guaranteed esignation of origin), IGP: Indicazione Geografica Protetta (Protected geographical indication).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

[7]

UL: Aceto **Balsamico** **(Tradizionale) (di Modena/ Reggio-Emilia)**

ESEMPI: 1. *L'aceto balsamico è un condimento molto popolare sulle tavole italiane.* 2. *L'antica acetaia, fondata nel XVIII secolo, produce un pregiato Aceto Balsamico di Modena.*

COLLOCAZIONI: *Aceto Balsamico di Modena IGP, Aceto Balsamico Tradizionale di Modena DOP, Aceto Balsamico Tradizionale di Reggio-Emilia DOP; glassa di aceto balsamico*

■ CULTURA:

L'Aceto Balsamico di Modena IGP è il prodotto ottenuto dai mosti appartenenti a sette varietà di uva del territorio emiliano. A questi mosti, parzialmente fermentati, cotti o concentrati, è prevista l'aggiunta di aceto vecchio di almeno 10 anni e minimo del 10% di aceto di vino. La quantità di mosto d'uva utilizzata non deve essere inferiore al 20%. È consentita l'aggiunta di caramello, per la stabilizzazione del colore, fino ad un massimo del 2% del volume del prodotto finito, ma non è ammessa l'aggiunta di altre sostanze oltre quelle già menzionate. L'acetificazione e l'affinamento avvengono in recipienti di legno pregiato quali quercia, rovere, castagno, gelso e ginepro, nell'arco di un periodo minimo di ٦٠ giorni. [...] (www.altroconsumo.it)

https://acetaiaestense.com/aceto-balsamico-di-modena

The Lexical Unit (in Italian: Unità Lessicale, UL) is already provided by the name of the product carrying the specification label. Being a proper name, it does not require a definition. However, in the Culture section (in Italian: 'Cultura'), a brief encyclopaedic description of Aceto Balsamico di Modena, followed by cross-references to external texts dealing with its origin production, properties, and its IGP or DOP status are crucial. *Aceto Balsamico* (*Tradizionale*) (*di Modena/Reggio-Emilia*), which is phraseological in nature, also builds complex recursive collocations (cf. Giacomini/DiMuccio-Failla/Lanzi 2020), some of which are further proper names. Images can also be useful to gain a first impression of the organoleptic characteristics of the product.

3) *to stagger* in the English PAD:

Motion verbs in English have several troponyms which correspond to very subtle meaning distinctions. This is not a culture-specific feature from a conceptual point of view. Rather, the cultural specificity concerns the extremely precise lexical differentiation, which may be missing in other languages. Among the troponyms of the verb to walk, WordNet lists, for example, *lollop*, *tap*, *stumble*, *sneek*, *swagger*, *scuffle*, *stagger* and many others. The motion verb *to stagger* can be presented as follows:

[7] https://upload.wikimedia.org/wikipedia/commons/thumb/4/4e/Balsamic_vinegar_%28drops%29.jpg/330px-Balsamic_vinegar_%28drops%29.jpg

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

[8]	LU: sb. **(ESP. a drunk) staggers around / in a ct. direction** DEFINITION: sb. (ESP. a drunk) walks very unsteadily around/ in a ct. direction EXAMPLES: 1. *The girl was walking unsteadily, too, stumbling and staggering from side to side.* 2. *He staggered a few feet then dropped to the ground.* COLLOCATIONS: *to stagger back/ backwards/ away/ forward/ sideways/ about; to stagger drunkenly*

Actions like *staggering* are best illustrated by videos. In this case, however, no additional Culture section is required.

4) *peer* in the English PAD:

[9]	LU: a **political party (ESP. Conservative/ Labour) peer in the House of Lords** DEFINITION: a member of the House of Lords belonging to a political party (ESP. Conservative/ Labour) EXAMPLES: 1. *The British government has continued the process of abolishing hereditary peers in the House of Lords.* 2. *A life peer cannot pass their title on to his or her children.* COLLOCATIONS: *hereditary/ Tory/ representative peer* ■ CULTURE: *peer* is a core element of the frame UK PARLIAMENT: UK Parliament is made up of three central elements: the **Monarchy** and two **Chambers**, namely the **House of Commons** and the **House of Lords / House of Peers**. **MPs** sit in the House of Commons. **(Members of the) Lords / peers** sit in the House of Lords. The main functions of the UK Parliament are **scrutinizing the work of the Government, legislation, debating, checking and approving Government spending.** [...]

Meanings and cultural implications of words indicating roles such as *professor*, *nurse*, *parent*, or *peer*, are best described in terms of a frame, which provides the user with situational knowledge (cf. Fillmore 2006). Within a frame, a word is embedded in a complex scenario made up of entities, events, states of affairs etc. A suitable frame for describing the role of a peer is the frame UK PARLIAMENT, in which core frame elements (bold characters) are presented together with their relations (underlined). The frame itself can be introduced by using a multimedia approach, e. g. by means of graphs, images and videos.

The above-mentioned examples show that different strategies should be taken into account for presenting culture-specific lexical items in a learner's dictionary. The PAD model provides, with its phraseological lexical units, first contextual syntactic-semantic information on a word and, at the same time, a key to first access cultural content: for instance, *sb. (ESP. a drunk) staggers around/ in a certain direction* matches the content of the video in a more

[8] https://www.youtube.com/watch?v=T09ufCyTtV0

[9] https://commons.wikimedia.org/wiki/File:House_of_Lords_2011.jpg

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/ Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

precise way than *to stagger* alone, while *political party* (ESP. *Conservative/Labour*) *peer in the House of Lords* anticipates the content of the frame description in a more explicit way than the noun *peer* alone.

A PAD aims to offer a holistic treatment of the lexicon starting from a phraseological view of language. The Culture section is a repository for cultural and/or situational knowledge required by a non-native user to fully understand a text while acquiring cultural competence (cf. Nied Curcio 2020). The same purpose is fulfilled by the media section of the microstructure, in which images and videos can be found. A further method for increasing efficiency in the presentation of culture-bound phenomena in a leaner's dictionary is the parallel treatment of similar words, which can be clustered in a homogeneous semantic field (e. g. motion verbs) or domain (e. g. design and architecture, food).

6. Conclusions

According to the broad notion of *cultural specificity* adopted in this paper, a large portion of the lexicon is affected by some kind of cultural influence. From a lexicographic point of view, culture-specific lexical items pose a considerable challenge to learner's dictionaries, which do not focus on a single target group but usually address speakers of many different languages.

To decide which entries need to be integrated with culture-related information is not an easy task for a lexicographer. Moreover, the amount and type of information required varies from word to word. As illustrated in the examples from the PADs, depending on the expected relation between concepts and lexicalisations in distinct cultures as well as on the type of entity (or: concept) we are dealing with, different strategies should be used to reach the same goal: to enable the user to fully understand the meaning and the cultural implications of a word or expression. A holistic, multimedia approach is desirable, because it offers multiple ways of communicating content:

- images (e. g. pictures and drawings) are particularly useful for illustrating concrete objects and states of affairs;
- videos are particularly useful for showing actions;

frames providing specific situational knowledge are particularly useful for describing events and roles.

There is no strict subdivision between these strategies, nor is their mutually exclusive application advisable. This has been shown in all examples of section 5. Besides prototypical images, videos and frames, further data can be added, also in the form of links to external sources.

The novelty of the PAD model lies in the fact that the first access to cultural content is granted by the phraseological structure of the Lexical Units, which provide the base syntactic-semantic context in which a word in a specific sense typically occurs. This phraseological pattern also determines the way in which culture-specific information is presented: in the same way that we do not define words in isolation but in their fundamental syntactic-semantic contexts, we do not present cultural information independently of the minimal context set by the Lexical Unit.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

References

Bödeker, B./Freese, K. (1987): Die Übersetzung von Realienbezeichnungen bei literarischen Texten: Eine Prototypologie. TextConText 2, pp. 137–165.

DiMuccio-Failla, P./Giacomini, L. (2017a): Designing an Italian learner's dictionary based on Sinclair's lexical units and Hanks's corpus pattern analysis. In: Proceedings of the Fifth eLex Conference Electronic Lexicography in the 21st Century. https://elex.link/elex2017/proceedings-download/ (last access: 28-06-2022).

DiMuccio-Failla, P./Giacomini, L. (2017b): Designing an Italian learner's dictionary with phraseological disambiguators. In: Mitkov, R. (ed.): Computational and Corpus-Based Phraseology. Second International Conference. Europhras 2017, LNAI 10596. Heidelberg, pp. 290–305.

Fillmore, C. J. (2006): Frame semantics. In: Cognitive linguistics: Basic readings 34, pp. 373–400.

Giacomini, L./DiMuccio-Failla, P. (2019): Investigating semi-automatic procedures in pattern-based lexicography. In: Proceedings of the eLex 2019 Conference. Electronic Lexicography in the 21st Century. Sintra, Portugal.

Giacomini, L./DiMuccio-Failla, P./Lanzi E. (2020): The interaction of argument structures and complex collocations: role and challenges for learner's lexicography. In: Proceedings of the XIX EuraLex International Congress, Alexandroupolis (GR). https://euralex2020.gr/wp-content/uploads/2020/11/EURALEX2020_ProceedingsBook-p285-293.pdf (last access: 28.6.2022).

Hanks, P. (2013): Lexical analysis: norms and exploitations. Cambridge, MA.

House, J. (2004): Culture-specific elements in translation. In: Kittel, H./Frank, A. P./Greiner, N./Hermans, T./Koller, W./Lambert,J./Paul, F. (eds.): Übersetzung: ein internationales Handbuch zur Übersetzungsforschung. Vol. 1. Berlin/New York, pp. 494–504.

Koller, W. (1979): Einführung in die Übersetzungswissenschaft. Heidelberg.

Kujamäki, P. (2004): Übersetzung von Realienbezeichnungen in literarischen Texten. In: Kittel, H./Frank, A. P./Greiner, N./Hermans, T./Koller, W./Lambert, J./Paul, F. (eds.): Übersetzung – Translation – Traduction. 1. Teilband (Vol. 1). Berlin/New York, pp. 920–925.

Kutz, W. (1977): Gedanken zur Realienproblematik (I). In: Fremdsprachen 1, pp. 254–259.

Markstein, E. (1998): Realia. In: Snell-Hornby, M. et al. (eds.): Handbuch translation. Tübingen, pp. 288–291.

Newmark, P. (1988): A textbook of translation. London.

Newmark, P. (1981): Approaches to translation. Vol. 1. Oxford.

Nied Curcio, M. (2020): Kulturell geprägte Wörter zwischen sprachlicher Äquivalenz und kultureller Kompetenz. Am Beispiel deutsch-italienischer Wörterbücher. In: Lexicographica 36, pp. 181–204.

Pamies, A. (2017): The concept of cultureme from a lexicographical point of view. In: Open Linguistics 3 (1), pp. 100–114.

Schreiber, M. (2007): Transfert culturel et procédés de traduction: l'example des realia. In: Kulessa von, R./Lombez, C. (eds.): De la traduction et des transferts culturels. Paris, pp. 185–194.

Sinclair, J. (2004): Trust the text: language, corpus and discourse. London.

Sobolev, L. N. (1952): Posobie po perevodu s russkogo jazyka na francuzskij (Moskau). Izdatel'stvo literatury na inostrannych jazykach.

Svensén, B. (2009): A handbook of lexicography: The theory and practice of dictionary-making. Kiribati.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Tomaszczyk, J. (2017): The culture-bound element in bilingual dictionaries. In: Hartmann, R. R. K. (ed.): LEXeter'83 Proceedings. Papers from the International Conference on Lexicography at Exeter, 9–12 September 1983 (reprint). Berlin/Boston, pp. 289–298.

Vlachov, S. I./Florin, S. (1980): Neperevodimoe v perevode. Mežd. Otnošenija.

Williams, J. (1990): The translation of culture-specific terms. In: Lebende Sprachen 35 (2), pp. 55 58. https://doi.org/https://doi.org/10.1515/les.1990.35.2.0.

Zhang, Y./Mi, H. (2020): Enhancing the role of culture-specific constructs in Chinese (-English) dictionaries for international learners. In: Lexicographica 36 (1), pp. 59–87.

Contact information

Laura Giacomini
University of Hildesheim
laura.giacomini@uni-hildesheim.de
University of Heidelberg
laura.giacomini@uni-heidelberg.de

Paolo DiMuccio-Failla
University of Hildesheim
muccio@uni-hildesheim.de

Patrizio De Martin Pinter
University of Heidelberg
patrizio.de_martin_pinter01@stud.uni-heidelberg.de

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Annette Klosa-Kückelhaus

LEXICOGRAPHY FOR SOCIETY AND WITH SOCIETY – COVID-19 AND DICTIONARIES

Abstract Not only professional lexicographers, but also people without a professional background in lexicography, have reacted to the increased need for information on new words or medical and epidemiological terms being used in the context of the COVID-19 pandemic. In this study, corona-related glossaries published on German news websites are presented, as well as different kinds of responses from professional lexicography. They are compared in terms of the amount of encyclopaedic information given and the methods of definition used. In this context, answers to corona-related words from a German question-answer platform are also presented and analyzed. Overall, these different reactions to a unique challenge shed light on the importance of lexicography for society and vice versa.

Keywords Lay-lexicography; professional lexicography; glossaries; general language dictionaries; neologism dictionaries

1. Introduction

One positive aspect of the COVID-19 pandemic – at least in the eyes of a lexicographer – is the fact that collecting and explaining all the vocabulary that was used in context of the pandemic has brought dictionaries and their editors some new-found publicity and increased interest in their work.[1] The neologism dictionary published at the Leibniz-Institute for the German Language (IDS) at Mannheim is one striking example of a dictionary that was impacted by the multi-layered interaction between lexicographers and society.[2] On 7 December 2020, the German news programme "Tagesschau" included a report on new corona-related words published as part of this dictionary at the IDS.[3] This was later watched by an Australian PhD student in linguistics living in Great Britain who tweeted on 21 February 2021 about the expansion of German vocabulary (cf. Fig. 1). Her tweet went viral and was picked up by the British media, e. g. "The Guardian" on 23 February 2021.[4] Then, the German media noticed the British interest in German corona neologisms and published articles on this phenomenon, e. g. the news magazine "Der Spiegel" on 2 March 2021.[5] Within the next couple of days, a total of 16 interviews were given by IDS lexicographers and 35 press articles were published – equivalent to the number of press articles on the neologism dictionary that normally appear in a single year. In addition, the number of times the dictionary was consulted increased considerably in the days following the English media coverage, from approximately 800 page retrievals and 300 visitors to the site on average per day to 18,000 page retrievals and 15,000 visitors (on 1 March 2021) and 17,700 page retrievals and

[1] For other linguistic studies on the effects of the pandemic on German see Klosa-Kückelhaus (ed.) (2021), Standke/Topalović (eds.) (2021), Weinert (2021) and Wengeler/Roth (eds.) (2020).

[2] For a discussion of "dictionaries in the public eye" see Curzan (ed.) (2021).

[3] See https://www.tagesschau.de/multimedia/sendung/tt-7945.html (last access: 15-03-2022).

[4] See https://www.theguardian.com/world/2021/feb/23/from-coronaangst-to-hamsteritis-the-new-german-words-inspired-by-covid (last access: 15-03-2022).

[5] See https://www.spiegel.de/wissenschaft/mensch/corona-wortschoepfungen-wo-wir-grossbritannien-in-der-pandemie-voraus-sind-a-7374e886-3556-4826-b987-6e894097ce0c (last access: 15-03-2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

13,500 visitors (on 2 March 2021). Finally, in the fortnight following the appearance of the press articles on the dictionary, almost 120 suggestions for the inclusion of new corona-related words were submitted by dictionary users via an online form,[6] compared to fewer than 100 suggestions that usually reach the editorial team annually.

Fig. 1: Tweet by Liz Hicks from 21 February 2021 on new corona-related words in German

Overall, this dictionary project has profited considerably from the widespread media coverage throughout the COVID-19 pandemic, not only through increased user numbers and new information on good candidates for inclusion into the list of corona-related neologisms, but also by being able to inform the public more generally about its work in newspapers and magazines, in podcasts, or in other media formats and by raising the awareness of the usefulness of lexicographic work for society in general.

But it is not only professional lexicography that has risen to the challenge of explaining the meaning of words from epidemiological, clinical, or medical contexts to bewildered speakers of many languages around the world or of collecting all these new words that have suddenly emerged. Non-lexicographical experts, like journalists or administrative staff in public institutions, have also helped to spread information about the meaning and usage of many corona-related expressions (for different German examples, see Möhrs 2021).

In this study, data on lay-lexicographic German glossaries compiled in the context of the COVID-19 pandemic is presented. Information about the number and types of entries (single words vs. multi-word units etc.) and the content of these publications is given. In a second step, these lay-lexicographic publications are compared to the responses of professional lexicographers to the increased need for word-related information during the COVID-19 pandemic, such as the updating of general language dictionaries or collection of corona-related neologisms, as well as the publication of articles, blogs etc. outside dictionaries on questions of lexical change during the COVID-19 pandemic and other topics.

The second part of this study focuses on comparing the methods of definition used by lay-lexicographers and professional lexicographers and on evaluating the amount of encyclopaedic information[7] given in their definitions of corona-related words. As far as the tech-

[6] See https://www.owid.de/wb/neo/mail.html (last access: 15-03-2022).

[7] Wiegand (1998, pp. 47 ff.) discusses how difficult it is to actually differentiate between what he calls "Sprachlexikographie" (i. e. dictionaries) and "Sachlexikographie" (i. e. encyclopaedias) and that very often lexicographers need to give at least some encyclopaedic information when defining word meanings.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

niques used in definitions are concerned, it is not only the definitions in professional dictionaries and glossaries produced by lay-lexicographers that are examined, but also the definitions found on websites like "gutefrage.net"[8] ('goodquestion.net') where users give answers to other users' questions, like "Was bedeutet Triage?" ('What is the meaning of triage?'). I will look into the form of the definitions (e.g. full sentence(s) vs phrase?) and whether or not they use the established lexicographic conventions for definitions that can be found in general monolingual dictionaries.

Overall, I hope to demonstrate through the data provided here how society is especially reliant on lexicographic traditions and lexicographic products in a context where speakers are confronted not only with an unprecedented situation in life, but also with the need to understand many terms previously unknown to them and a large number of new lexical items emerging in a short time span.

2. Lexicographic responses to users' needs in the COVID-19 pandemic

From the beginning of the COVID-19 pandemic, new words emerged in general language (foremost, the names of the new virus, *SARS-CoV-2*, and of the disease it causes, *COVID-19*), and specific terms taken from the medical or epidemiological context became part of everyday language. Speakers all over the world had to adapt quickly, sometimes needing help in understanding certain words or phrases, in correctly differentiating between them (e.g. *pandemic* – *epidemic*), and in using them correctly (grammatically, orthographically, etc.). This need was felt and reacted to by lexicographers around the world,[9] but also, for example, by people who are not lexicographic experts, like journalists:

In Germany, the first glossary of corona-related words and phrases was published in early March 2020 by the newspaper the "Süddeutsche Zeitung", even before the WHO officially declared the SARS-CoV-2 pandemic on 11 March 2020 (see Möhrs 2020). Some of the reasons for their publication were given, as shown in (1):

(1) Inzwischen kennen wir viele Begriffe rund um das Coronavirus und die Krankheit Covid-19. Viele Begriffe sind aber schwer zu verstehen und erklärungsbedürftig, zudem kommen ständig neue Fremdwörter hinzu. Die folgende Liste soll einige wichtige Begriffe […] erklären […].[10] ('By now we are familiar with many terms surrounding the coronavirus and the disease Covid-19. However, many terms are difficult to understand and need explanation; moreover, new foreign words are being added all the time. The following list is intended to explain some important terms […].')

Lexicographers from the "Digitales Wörterbuch der Deutschen Sprache" (DWDS) at the Berlin Academy of Sciences published a first version of their "Corona Glossar" in the middle

[8] See https://www.gutefrage.net/ (last access: 15-03-2022).

[9] See the talks given at the "3rd Globalex Workshop on Lexicography and Neology – Focus on Corona-related Neologisms 2021, @AUSTRALEX 2021", https://globalex2021.globalex.link/ (last access: 15-03-2022).

[10] See https://www.nzz.ch/visuals/coronavirus-diese-20-begriffe-rund-um-covid-19-muessen-sie-kennen-ld.1553235 (last access: 15-03-2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

of March 2020, which has been updated continuously since then.[11] In (2) they explain their actions:

(2) Die COVID-19-Pandemie löst weltweit zahlreiche Prozesse des Wandels aus, die ihren Niederschlag auch in der Sprache finden. […] Die Redaktion des DWDS sieht es als wichtige Aufgabe an, diese Veränderungen zeitnah zu dokumentieren und in die Wörterbucheinträge des DWDS zu integrieren.[12] ('The COVID-19 pandemic triggers numerous processes of change worldwide, which also find expression in language. [...] The editorial staff of the DWDS considers it an important task to document these changes in a timely manner and to integrate them into the dictionary entries of the DWDS.')

In the following section, more information on the reaction of journalists (acting as lay-lexicographers) and professional lexicographers to the challenge is presented and compared.

2.1 Journalistic glossaries with corona-related vocabulary

For this study, data from ten glossaries published on news sites from Austria, Germany, and Switzerland was collected and evaluated (cf. Table 1).[13] These lists were chosen because they address the general public, which is also the aim of general language dictionaries (see 2.2). In addition, general language and neologism dictionaries of German are mostly based on newspaper and magazine corpora. Thus, glossaries and word lists published by newspapers and magazines seemed the most obvious option for a comparison between professional and lay-lexicographic explanations of corona-related words.

Other glossaries specifically addressing children (e. g. "Corona-Lexikon: Sprichst du Coronisch?" by "BR Kinder"[14]) or written in simple language (e. g. "Lexikon Corona leichte Sprache" by "Task Force Leichte Sprache/Anne Leichtfuß"[15]) were excluded, as well as some stemming from a medical context (e. g., "Corona-Fachbegriffe kurz erklärt" by "Medizinische Hochschule Hannover"[16] [Hannover Medical School]) or published by the public authorities (e. g. "Glossar Coronavirus" by "Bayerisches Landesamt für Gesundheit und Lebensmittelsicherheit"[17] [Bavarian State Office for Health and Food Safety]). All of these are not comparable to glossaries from newspapers and magazines because of their different target groups and their terminological or official character, all of which might influence the selection of head words, the methods of definition, and the amount of encyclopaedic information given.

[11] Specific terminological dictionaries on corona-related terms have also been published quickly, e. g. "Corona-Terminologie Deutsch, Englisch, Französisch, Niederländisch, Polnisch, Russisch, Spanisch" by Bundessprachenamt (Federal Language Bureau) on 24 April 2020, see https://app.coreon.com/5ea2adb797e1040100eb7ff3/concepts/5ec2ac4d238bd30039949e9d (last access: 15-03-2022).

[12] See https://www.dwds.de/themenglossar/Corona (last access: 15-03-2022).

[13] I would like to thank Sara-Marie Weinkopf (IDS Mannheim) for her support in the compilation of all data.

[14] See https://www.br.de/kinder/corona-lexikon-sprichst-du-coronisch-100.html (last access: 15-03-2022).

[15] See https://corona-leichte-sprache.de/lexikon/ (last access: 15-03-2022).

[16] See https://corona.mhh.de/corona-glossar (last access: 15-03-2022).

[17] See https://www.lgl.bayern.de/gesundheit/infektionsschutz/infektionskrankheiten_a_z/coronavirus/covid_glossar.html (last access: 15-03-2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Newspaper/ Newsmagazine and title of glossary	Number of entries	Order of head words
Deutschlandfunk: "Covid-19-Glossar/Die wichtigsten Begriffe zur Coronavirus-Pandemie"[18]	28	alphabetical
Die Rheinpfalz: "Virus-Lexikon: Corona von A bis Z"[19]	26	
Der Spiegel: "Zentrale Begriffe der Coronakrise – und was sie bedeuten"[20]	16	
Tagesschau: "Die wichtigsten Corona-Begriffe"[21]	15	
T-online: "Alle wichtigen Begriffe rund um Corona"[22]	21	
Bild der Frau: "Coronavirus-Glossar: Diese Begriffe sollten Sie kennen"[23]	23	not alphabetical[24]
GEO: "Corona-Glossar: Diese Begriffe sollten Sie jetzt kennen"[25]	18	
Iserlohner Kreisanzeiger: "Coronavirus-Glossar – Begriffe, die Sie jetzt kennen sollten"[26]	11	
Neue Zürcher Zeitung: "Das Glossar zum Coronavirus – Corona-Impfung, Varianten, Übersterblichkeit und andere Begriffe, die Sie kennen müssen"[27]	31	
Profil: "Das Coronavirus-Glossar – Wichtige Begriffe aus dem Corona-Krisen-Vokabular einfach erklärt"[28]	7	
Total	196	

Table 1: Overview of glossaries on corona-related words and phrases from German news websites

[18] See https://www.deutschlandfunk.de/covid-19-glossar-die-wichtigsten-begriffe-zur-coronavirus-100.html (last access: 15-03-2022).

[19] See https://rp-online.de/panorama/coronavirus/coronavirus-lexikon-das-bedeuten-triage-inkubation-mortalitaet-und-co_aid-49588865 (last access: 15-03-2022).

[20] See https://www.spiegel.de/politik/deutschland/corona-pandemie-glossar-zentrale-begriffe-in-der-corona-krise-und-was-sie-bedeuten-a-8f0d15a9-c7ed-4265-bf21-f8d7f6836292 (last access: 15-03-2022).

[21] See https://www.tagesschau.de/inland/corona-pandemie-glossar-101.html (last access: 15-03-2022).

[22] See https://www.t-online.de/gesundheit/krankheiten-symptome/id_87524194/coronavirus-glossar-zu-wichtigen-begriffen-rund-um-corona.html (last access: 15-03-2022).

[23] See https://www.bildderfrau.de/gesundheit/krankheiten/article228837093/Coronavirus-Glossar-Begriffe-erklaert.html (last access: 15-03-2022).

[24] For details, see below.

[25] See https://www.geo.de/wissen/gesundheit/22805-rtkl-kurz-erklaert-corona-glossar-diese-begriffe-sollten-sie-jetzt-kennen (last access: 15-03-2022).

[26] See https://www.ikz-online.de/panorama/diese-begriffe-rund-um-das-coronavirus-sollten-sie-kennen-id228830547.html (last access: 15-03-2022).

[27] See https://www.deutschlandfunk.de/covid-19-glossar-die-wichtigsten-begriffe-zur-coronavirus-100.html (last access: 15-03-2022).

[28] See https://www.profil.at/wissenschaft/das-coronavirus-glossar/400877954 (last access: 15-03-2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Each chosen glossary explains up to approximately thirty corona-related head words[29]; as a whole, these lists comprise 112 different words or phrases. Table 2 gives information on how many head words or phrases are included in more than one of the glossaries (*COVID-19*, for example, is presented in all of the ten glossaries and *SARS-CoV-2* in eight of them) and differentiates between single word entries (e. g. *asymptomatisch* 'asymptomatic') and multi word entries (e. g. *aktive Fälle* 'active cases'). Most of the nouns in the glossaries are given in their singular form (e. g. *Boosterimpfung* 'booster vaccination'), but not all (e. g. *Coronaviren* 'corona viruses'), and a high proportion of the headwords stem from medical or epidemiological contexts (e. g. *Mortalitätsrate* 'mortality rate', *Containment* 'containment').

Some of the glossaries studied here not only contain encyclopaedic information or definitions, but also provide illustrations, e. g. the *Neue Zürcher Zeitung* from Switzerland shows a graph depicting a human figure being injected with a syringe in its upper arm alongside the entry *Impfung gegen das Coronavirus* ('vaccination against the corona virus'). Only in one case is a hyperlink to another website given (from the head word *Desinfektion/Hygiene* in "Der Spiegel",[30] a German news magazine). None of the glossaries presents information on the syllabification, pronunciation, or grammar of the head words.

Numbers of entries	contained in one glossary	contained in up to 10 glossaries
	71.5 %	28.5%
Types of entries	single word entries	multi word entries
	83 %	17 %
Grammatical features	singular	plural[31]
	83.5 %	13%
General or specialized contexts	epidemiological or medical context	other contexts
	83%	17%

Table 2: Different head word types in corona glossaries from German new websites (N = 112)

While most of the head words are given in their typical German dictionary form (nouns: nominative singular, adjectives: without inflection, verbs: infinitive), something else is noteworthy: several entries in the glossaries explain more than one word, e. g. entries like *Ausgangssperren/Ausgangsbeschränkungen* ('curfews/restrictions on going out') in "Der Spiegel" or *Coronavirus/SARS-CoV-2/Covid-19* in the glossary of the Austrian newspaper "Profil". Unlike in professional dictionaries, these entries on (near) synonyms economize on space by presenting only one explanation for two or more expressions. Another difference to most general language dictionaries is that half of the glossaries evaluated for this study do not list the entries in alphabetical order, but in thematic order (e. g. in the "Neue Zürcher Zeitung") or in no recognizable order at all (e. g. in "GEO") (cf. Table 1).

[29] There are several German glossaries and word lists on corona-related words containing more entries than 30 (for more details, see Möhrs 2021). These were excluded here to restrict the material to a manageable size. Further studies may incorporate these other materials.

[30] This entry links to the web page of the "Bundeszentrale für gesundheitliche Aufklärung" (Federal Center for Health Education).

[31] The remainder do not allow for a distinction between singular or plural, e. g. verbal phrases.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

2.2 Responses from professional lexicography

In this section, three different ways of reacting to the increased demand for lexicographic information on (new) words used in the context of the COVID-19 pandemic are discussed: firstly, updating general language dictionaries in relation to corona-related entries; secondly, collecting and publishing neologisms and older words and phrases from the pandemic in specialized lists; and finally publishing articles or blogs etc. outside dictionaries on questions of lexical change during the pandemic.

Figure 2 shows the entry *Variantengebiet* in the DWDS,[32] which was updated on 24 August 2021. The older sense of 'region where off-piste skiing is possible' and the corona-related sense 'country or part of the country where a (dangerous, more infectious) variant of a pathogen is dominant (which is why special measures must be taken)' are given. Probably most of the general language dictionaries around the world (especially those published online) have been updated throughout the COVID-19 pandemic in a similar way.

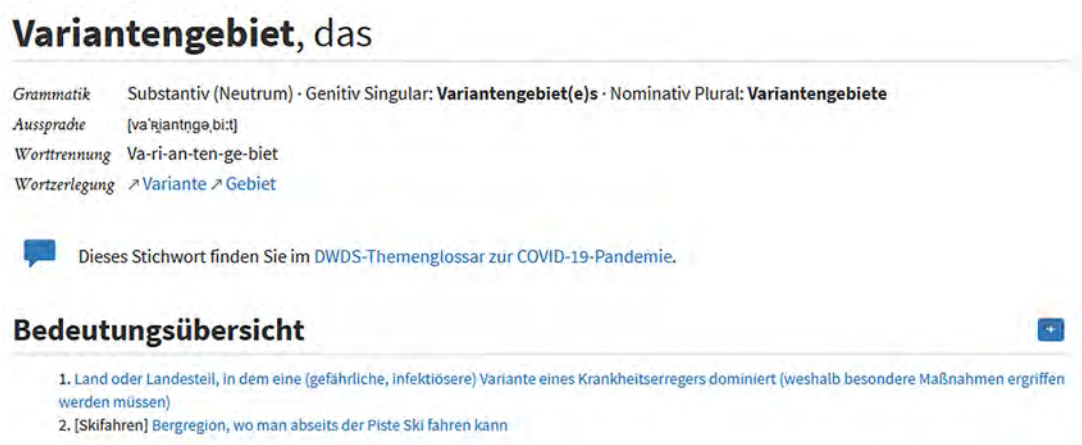


Fig. 2: Entry *Variantengebiet* in the DWDS

In Figure 3, an extract with the entry *teamsen* ('to communicate by using the video conference system Teams®') from the glossary "Neuer Wortschatz rund um die Coronapandemie"[33] ('New Vocabulary surrounding the corona pandemic'), published as part of the German neologism dictionary at IDS Mannheim, illustrates how dictionary projects for German reacted quickly to language change in the pandemic by compiling lists relating to this specific area of vocabulary. Here, an existing format of entries, currently used to present a list of new words still being monitored for possible inclusion in the dictionary,[34] was adapted to present lexicographic information on new words and phrases from the pandemic.

[32] See https://www.dwds.de/wb/Variantengebiet (last access: 15-03-2022).

[33] See https://www.owid.de/docs/neo/listen/corona.jsp#teamsen (last access: 15-03-2022).

[34] See https://www.owid.de/docs/neo/listen/monitor.jsp (last access: 15-03-2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

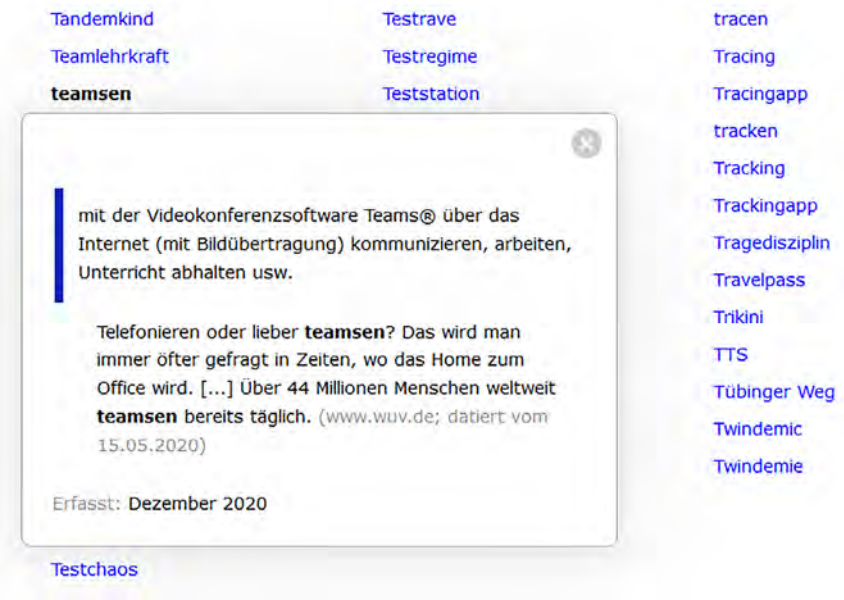


Fig. 3: Extract from the "Neuer Wortschatz rund um die Coronapandemie", showing the entry *teamsen*

Several dictionary projects also published informative articles on the expansion of vocabulary during the pandemic or on the specific challenges involved in using corona-related words, for example as texts outside the dictionaries. In Figure 4, an article on the distinction between *Epidemie* and *Pandemie* from "Duden online"[35] is shown. In the introductory paragraph of this text, the Duden lexicographers explain their intention: "In Sachen Coronavirus werden wir mit einer Fülle von medizinischen Informationen konfrontiert, die oft eine

Fig. 4: Text on the difference between *Epidemie* and *Pandemie* in "Duden online"

[35] See https://www.duden.de/sprachwissen/sprachratgeber/Epidemie-Pandemie (last access: 15-03-2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

gewisse Ratlosigkeit hinterlassen. Damit Sie zumindest sprachlich nicht ratlos sind, greifen wir ein paar passende Themen auf: […]" ('When it comes to coronavirus, we are confronted with a wealth of medical information that often leaves us with a sense of helplessness. To ensure that you are not left helpless, at least linguistically, we will take up a few appropriate topics: […]').

Overall, professional German lexicography did not change its traditional ways of listing words: new (predominantly single word) entries are given in their typical dictionary form (see 2.1), but they do not explain more than one word or more than one multi-word unit in a single article, as is the case in quite a number of entries in the newspaper glossaries (see 2.1). Neither do entries on new words differ from articles on older words (with new meanings) in their microstructure or data presentation (online). In the three German online dictionaries discussed here, new corona-related entries have also been fully integrated into the search options. Using external texts to give more information on specific groups of words is not a new feature in (online) lexicography either. More detailed studies are needed when it comes to the extent to which new terms are integrated into the dictionaries shown here compared to new words or meanings in general language. In the following section, differences and similarities between lexicographic, journalistic, and laypersons' methods of definition in corona-related entries are examined.

3. Methods of definition

A central part of the micro-structure of most dictionaries is the lexicographic definition (Wiegand 1989, p. 531), given either as a phrase (mostly in general language dictionaries) or in a full sentence (mostly in learner lexicography).[36] There are two established lexicographic traditions for definitions in general monolingual dictionaries: giving a synonym (Haß-Zumkehr 2001, p. 28) or naming genus proximum and differentia specifica (ibid., p. 29). In addition, examples are used to illustrate meaning. In the following sections, the types of explanations given for head words in journalistic corona-related glossaries and in vocabulary-related questions and answers given by laypersons on the website "gutefrage.net" are analyzed. I will also evaluate the amount of encyclopaedic information contained in these definitions. Finally, the results are compared to the definitions provided in professionally compiled dictionary entries for corona-related words and phrases.

3.1. Definitions in lay-lexicographic glossaries with corona-related vocabulary

General data on ten glossaries with corona-related vocabulary published by news websites in Germany, Austria, and Switzerland is presented above (see 2.1). In this section, the explanations given in all 196 entries for all of the 112 words or phrases in the glossaries are analyzed to find out which defining techniques they use and how much encyclopaedic information they contain. All definitions were coded according to the following criteria:

- formal type of definition (full sentence and/or phrase),
- purely encyclopaedic explanation or hyperlinks to other websites with encyclopaedic information,

[36] See Hanks (2016) and Wiegand (1989), for more information on various approaches to writing lexicographic definitions.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

- lexicographic definitions (definitions with genus proximum and differentia specifica or synonyms),
- lexicographic definitions containing other lexicographic information (etymology, domain).

Table 3 summarizes the numbers.

Formal type of definition (N= 196)[37]	full sentence	phrase
	94%	14%
Purely encyclopaedic explanations vs. lexicographic definitions	encyclopaedic explanation	lexicographic definition
	24%	76%
Types of definitions (N= 144)	definition with genus proximum and differentia specifica	synonym(s)
	85%	15%
Other lexicographic information (N = 36)	etymology	domain
	75%	25%

Table 3: Different formal types of explanations and types of information in corona-related glossaries from German new websites

In over 90 cases, lexicographic definitions are supplemented by encyclopaedic information, e.g. in the explanation for *Inkubationszeit* ('incubation period') in (3) from the glossary in "Bild der Frau". Overall, 71% of all entries in the glossaries discussed here contain encyclopaedic information, sometimes long paragraphs with many sentences. In contrast, almost none of the glossaries use illustrations to convey information or offer hyperlinks to other online reference works (see 2.1).

(3) Dies ist die Zeitspanne zwischen Infizierung und Ausbruch der Krankheit mit Symptomen. Je nach Erkrankung kann die Inkubationszeit zwischen einigen Stunden und sogar einigen Jahren (etwa bei der von HIV ausgelösten Krankheit Aids) liegen. Im Schnitt liegt die Inkubationszeit bei Covid-19 bei fünf Tagen, kann aber auch bis zu 14 Tage dauern. Manche Infizierte verspüren hingegen sogar gar keine Symptome.[38] ('This is the period of time between infection and the onset of the disease with symptoms. Depending on the disease, the incubation period can range from a few hours to even several years (for example, in the case of AIDS, a disease caused by HIV). On average, the incubation period for Covid-19 is five days, but it can last up to 14 days. Some infected persons, by contrast, experience no symptoms at all.').

In (3), the first sentence gives the lexicographic definition in the classic form of genus proximum (*Zeitspanne* ['period of time']) and differentia specifica as a prepositional phrase (*zwischen Infizierung und Ausbruch der Krankheit mit Symptomen* ['between infection and the onset of the disease with symptoms']). The next three sentences present encyclopaedic information. Example (4) with information on the word *Abstrich* ('pap smear') illustrates a purely encyclopaedic entry in one of the glossaries ("Rheinlandpfalz Online").

(4) Damit der Test durchgeführt werden kann, benötigt das Labor Sekret aus dem Hals und der Nase, manchmal auch aus der Lunge. Dazu wird ein Wattestäbchen tief eingeführt und dann ins Labor geschickt. Bluttests auf Antikörper gibt es zwar auch schon, sie sind aber sehr un-

[37] In some cases, full sentences and phrases are combined in one entry.

[38] See https://www.bildderfrau.de/gesundheit/krankheiten/article228837093/Coronavirus-Glossar-Begriffe-erklaert.html (last access: 15-03-2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

sicher; außerdem könnten sie nur das Vorhandensein von (möglicherweise älteren) Antikörpern bestätigen und wären deshalb kein sicherer Hinweis auf eine akute Infektion.[39] ('In order for the test to be performed, the laboratory needs secretions from the throat and nose, sometimes also from the lungs. For this purpose, a cotton swab is inserted deeply and then sent to the laboratory. Blood tests for antibodies also already exist, but they are very unreliable; moreover, they would only be able to confirm the presence of (possibly older) antibodies and would therefore not be a reliable indication of an acute infection.')

Etymological information or information on synonyms is mostly included in the definitions as shown in examples (5) (entry *FFP* in "Bild der Frau") and (6) (entry *Mortalitätsrate* ['mortaliy rate'] in "T-online"). In (5), the first sentence explains that *FFP* is an abbreviation of English *filtering face piece*. In (6), the synonym *Sterberate* is given as an insertion in dashes.

(5) FFP ist die Abkürzung für das englisch „filtering face piece", also filternde Gesichtsmaske, zu Deutsch genauer „partikelfilternde Halbmaske". Hier handelt es sich tatsächlich um eine Atemschutzmaske. Sie schützt den Träger je nach Stärke mehr oder weniger vor dem Einatmen von kleinen und kleinsten (FFP3) Partikeln, die die Gesundheit schädigen könnten.[40] ('FFP is the abbreviation for "filtering face piece" or, more precisely, "particle-filtering half mask". This is actually a respiratory protection mask. Depending on its strength, it protects the wearer to a greater or lesser extent from inhaling small and very small (FFP3) particles that could be harmful to health.')

(6) Die Mortalitätsrate – auch Sterberate – bezeichnet die Anzahl der Todesfälle in einem bestimmten Zeitraum bezogen auf 1.000 Individuen einer Population. Als Zeitraum wird in der Regel ein Jahr definiert. Die Sterblichkeitsrate bei einer Infektion mit dem Coronavirus liegt laut Weltgesundheitsorganisation (WHO) bei bis zu drei Prozent (Vergleich Grippe < 1 Prozent).[41] ('The mortality rate – also the death rate – is the number of deaths in a given period per 1,000 individuals in a population. The period is usually defined as one year. According to the World Health Organization [WHO], the mortality rate for infection with the coronavirus is up to three percent [comparison: influenza < 1 percent].')

There is clearly a focus on encyclopaedic information in the corona-related glossaries published by journalists on German news websites analyzed here. In contrast to dictionary entries, the glossary entries are also almost exclusively written in full sentences or combine phrases and full sentences, while there are hardly any cases where the explanation is only given in a single phrase resembling a typical dictionary entry. Some kind of lexicographic definition of meaning is found in roughly two thirds of the entries, and other information typical for dictionaries (i. e. information on etymology or domain) is added in roughly one fifth of all entries.

When defining the meaning of the head word, the authors of the glossaries use only two traditional strategies of explaining meaning: defining by naming the genus proximum and differentia specifica or by naming synonyms (the latter, however, in only 15% of the cases). There are no cases where examples are used to illustrate meaning. Nevertheless, the lay-lexicographers at work on these glossaries are building on traditional ways of defining lexical meaning (as in other contexts where laypersons define words; see Klosa/Stähr 2019 and 2020).

[39] See https://rp-online.de/panorama/coronavirus/coronavirus-lexikon-das-bedeuten-triage-inkubation-mortalitaet-und-co_aid-49588865 (last access: 15-03-2022).

[40] See https://www.bildderfrau.de/gesundheit/krankheiten/article228837093/Coronavirus-Glossar-Begriffe-erklaert.html (last access: 15-03-2022).

[41] See https://www.t-online.de/gesundheit/krankheiten-symptome/id_87524194/coronavirus-glossar-zu-wichtigen-begriffen-rund-um-corona.html (last access: 15-03-2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

3.2 Definitions on "gutefrage.net"

The website "gutefrage.net" offers the opportunity to anybody to ask any kind of question and is the largest question-answer platform in German, with 1.9 billion active users and up to 30,000 answers given a day.[42] More than 100 questions concern language, which is one of the most popular topics. For this study, "gutefrage.net" was searched for all 112 entries found in the journalistic glossaries of corona-related words and expressions (see 2.1). Only seventeen of these were contained in questions and answers on this platform[43] (cf. Table 4).

Word or phrase	**number of questions**	**number of answers**
Aerosol ('aerosol')	1	3
COVID-19	1	1
Epidemie ('epidemic')	3	8
exponentiel[les Wachstum] ('exponential [growth]')	2	5
Inkubationszeit ('incubation period')	1	1
Inzidenz ('incidences')	1	1
Latenzzeit ('latency period')	1	1
Letaliät ('lethality')	2	6
linear[es Wachstum] ('linear [growth]')	2	4
Lockdown ('lockdown')	2	8
Mortalität ('mortality')	2	6
Pandemie ('pandemic')	3	11
Schmierinfektion ('smear infection')	1	1
Shutdown ('shutdown')	2	8
Sterberate ('mortality rate')	1	3
Triage ('triage')	3	16
Tröpfcheninfektion ('droplet infection')	2	3
Total	30	86

Table 4: Corona-related words and phrases explained on the question-answer-platform "gutefrage.net"

All the words and phrases that users wanted to obtain information about are terms from the medical or epidemiological context and hardly any are new words or expressions that came to prominence during the corona-pandemic. Three different types of questions can be found (in descending order of frequency):

- difference between A and B (e.g. *Was ist der Unterschied zwischen Pandemie und Epidemie?*)
- meaning of X (e.g. *Was heißt Triage?*, *Was bedeutet Shutdown?*, cf. Figure 4)
- definition of X (e.g. *Was ist Triage?*)

[42] See information on https://www.gutefrage.net/ (last access: 15-03-2022).

[43] And not all questions were posted and answered during the COVID-19 pandemic, i.e. from January 2020 to March 2022, but some also before January 2020.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

All the definitions found in the answers were coded according to the following criteria:
- answers containing no definitions (eleven answers, which consist, for example, only of a hyperlink to other websites, e. g. Wikipedia);
- formal type of definition (full sentence or phrase);
- definitions with encyclopaedic information or hyperlinks to other websites with encyclopaedic information;
- definitions presenting lexical information (definitions with genus proximum and differentia specifica or synonyms, examples);
- answers with other lexicographic information (etymology, domain).

Figure 5 shows a typical example of a definition in sentence form, listing synonyms as well as presenting encyclopaedic information.[44]

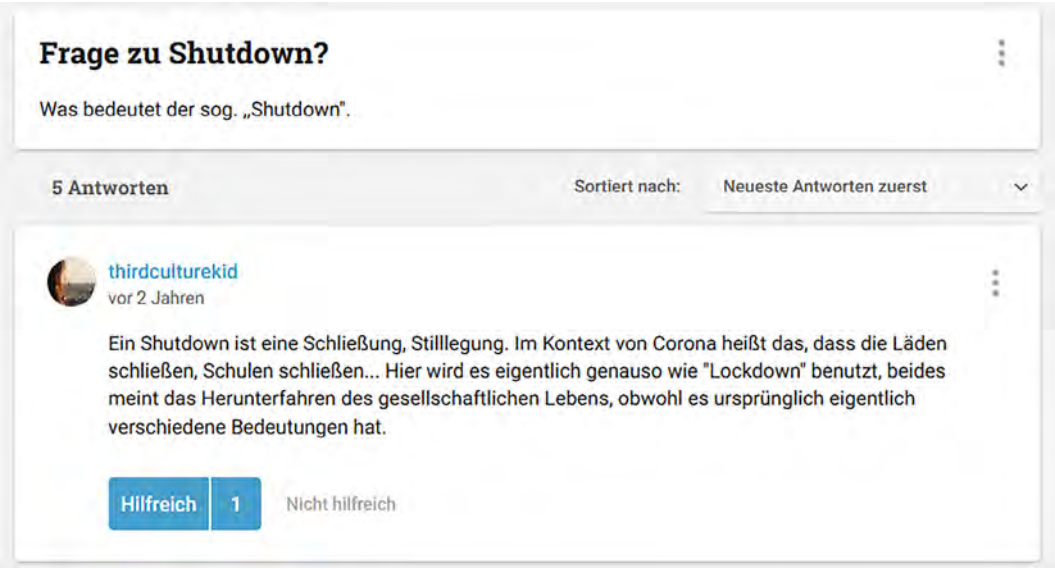


Fig. 5: Question on the meaning of *Shutdown* on gutefrage.net

The numbers in Table 5 illustrate that the answers given to questions in "gutefrage.net" evidently do not aim to be lexicographic definitions. They are predominantly written in full sentences and contain encyclopaedic information in over half of the cases. Purely lexical information is presented in less than half of the answers analyzed here, and other information typical for dictionaries is only added in fewer than 20% of the answers.

When defining the words that users asked for, three traditional strategies of explaining meaning are used to an almost identical degree: defining by naming the genus proximum and differentia specifica, by giving synonyms, and by showing examples of usage. Here (as in other contexts where lay people define words, see Klosa/Stähr 2019 and 2020), the users of the platform build on their knowledge of defining techniques possibly gained from teaching contexts or experience of dictionary usage.

[44] See: https://www.gutefrage.net/frage/frage-zu-shoutdown (last access: 15-03-2022). Translation of the answer: "A shutdown is a closure, shutting down. In the context of corona, it means that stores are closing, schools are closing [...] Here it's actually used the same way as "lockdown", both meaning the shutting down of social life, although originally it actually has different meanings."

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Formal type of definition (N= 57)	full sentence	phrase	
	88%	12%	
Encyclopaedic information (N = 46)	explicit encyclopaedic information	hyperlinks to other websites with encyclopaedic information	
	70%	30%	
Lexical information (N = 36)	definition with genus proximum and differentia specifica	synonym(s)	example(s)
	34%	30%	36%
Other lexicographic information (N = 10)	etymology	domain	
	40%	60%	

Table 5: Different formal types of explanations and information in definitions on the question-answer-platform "gutefrage.net"

It is noteworthy that over half of the questions and answers analyzed cover two words. The newspaper glossaries presented in section 2.1 also contain some entries on two words or more, e. g. on *Schmierinfektion* ('smear infection') and *Kontaktinfektion* ('contact infection'). This might be an indication for lexicography that users not only try to find information on single words, but also on near synonyms or word pairs stemming from the same domain.[45]

3.3 Comparison to professional lexicography

As shown in section 2.2, German (online) dictionaries have been updated either by adding new corona-related entries or new meanings or by publishing word lists with vocabulary used in the context of the COVID-19 pandemic. In both scenarios, the dictionaries have not changed the ways they define head words. The example *Inkubationszeit* 'incubation period' from the DWDS (7) and the head word *Shutdown* ('shut-down') from the neologism dictionary published at the Leibniz-Institut für Deutsche Sprache (IDS) Mannheim (8) show that the definitions consist of phrases in which the genus proximum (*Zeit* or *Zeitraum* 'time (span)') is named and specified. There is no encyclopaedic information. In (7), the domain "medicine" is given. Both entries also contain examples to explain meaning and usage.

(7) [Medizin] Zeit zwischen der Ansteckung mit einem Krankheitserreger und dem Auftreten der ersten Krankheitssymptome ('[medicine] time span between infection with a pathogen and the appearance of the first symptoms of the disease')[46]

(8) Zeitraum, in dem fast alle wirtschaftlichen und gesellschaftlichen Aktivitäten auf politische Anordnung hin stillgelegt sind (z.B. zur Eindämmung einer Seuche) ('period of time during which almost all economic and social activities are shut down by political order (e. g. to contain a disease)')[47]

[45] One notable exception is the German "Paronymwörterbuch", a corpus-based dictionary on easily confusable words published by the IDS Mannheim (see https://www.owid.de/parowb/, last access: 15-03-2022). Here, word pairs or triplets are explained contrastively in one dictionary entry. For more information, see Storjohann (2018).

[46] See https://www.dwds.de/wb/Inkubationszeit (last access: 15-03-2022).

[47] See https://www.owid.de/docs/neo/listen/corona.jsp#shutdown (last access: 15-03-2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Quite obviously, the authors of journalistic glossaries on corona-related vocabulary as well as the authors of answers to word-related questions posted on the question-answer platform "gutefrage.net" focus more on presenting encyclopaedic information than the professional lexicographers involved in German dictionary projects. When explaining word meaning, lay-authors prefer full sentences, but also fall back on traditional ways of defining words, while professional German lexicographers do not deviate from the conventional forms of definitions and the typical inventory of lexicographic information in their dictionaries.

4. Conclusion

To help users of German to understand and correctly use new corona-related words and the existing phrases and medical, epidemiological, political etc. terms being used more often in general language throughout the COVID-19 pandemic, it is not only professional lexicographers who have reacted quickly, but the need to explain these words and phrases was felt in other parts of society as well. Thus, for example, journalists published glossaries and word-related questions were being answered on question-and-answer platforms. This study found differences concerning the explanation of word meaning and the amount of encyclopaedic information given in professional dictionaries, on the one hand, and in lay-lexicographic formats, on the other. An interesting question to be asked is how satisfied users are regarding professional lexicographic information in contrast to lay-lexicographic information (within the context of corona-related words). This is not part of this study but could be fruitfully pursued in future research.

Overall, general interest in (new) words or phrases and their meaning and usage in German was strong, as was interest in the ways that dictionaries reacted to this challenge. This was shown in one striking example (see section 1) where the British media picked up on a tweet concerning interesting new German corona-related words, leading to a drastic increase of page views in one German online dictionary in a short period of time. There is an interest in lexicography and lexicographical practice in society that manifested itself most obviously in the publication of glossaries etc. by non-lexicographers, but also in increased consultation of (online) dictionaries and in a higher demand for interviews on vocabulary developments in the COVID-19 crises. Both sides may profit: society, because reliable dictionaries reacting quickly to language change supports effective communication; and lexicography, because of the higher awareness of the usefulness of dictionaries, hopefully resulting in both strong sales of dictionaries and continued public funding for academic dictionary projects, as well as a general acknowledgement of the importance of lexicographic work.

References

Curzan, A. (ed.) (2021): Forum "Dictionaries in the Public Eye". In: Dictionaries: The Journal of the Dictionary Society of North America 42 (1), pp. 211–257.

Haß-Zumkehr, U. (2001): Deutsche Wörterbücher. Brennpunkt von Sprach- und Kulturgeschichte. Berlin/New York.

Hanks, P. (2016): Definition. In: Durkin, P. (ed.): The Oxford handbook of lexicography. Oxford, pp. 94–122.

Klosa-Kückelhaus, A. (ed.) (2021): Sprache in der Coronakrise. Dynamischer Wandel in Lexikon und Kommunikation. Mannheim.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Klosa-Kückelhaus, A./ Stähr, L. (2019): Lexikographie an der Wand: Wörterbuchartikel als Wandtattoos und Poster. In: Sprachreport 2019 (4), pp. 26–35.

Klosa-Kückelhaus, A./Stähr, L. (2020): T-shirt lexicography. In: Dictionaries: The Journal of the Dictionary Society of North America 41 (2), pp. 93–113.

Möhrs, Chr. (2021): Grübelst du noch oder weißt du es schon? Glossare erklären Corona-Schlüsselbegriffe. In: Klosa-Kückelhaus, A. (ed.): Sprache in der Coronakrise. Dynamischer Wandel in Lexikon und Kommunikation. Mannheim, pp. 19–27.

Standke, J./Topalović, E. (eds.) (2021): In Krisen erzählen – von Krisen erzählen. Sprachliche, literarische und mediale Dimensionen. (= Mitteilungen des Deutschen Germanistenverbandes 68, Issue 2). Göttingen.

Storjohann, Petra (2018): Commonly confused words in contrastive and dynamic dictionary entries. In: Čibej, J./Gorjanc, V./Kosem, I./Krek, S. (eds.): Proceedings of the XVIII EURALEX International Congress. Lexicography in Global Contexts, 17–21 July, Ljubljana. Ljubljana, pp. 187–197.

Weinert, M. (2021): Krisensprache – Sprachkrise – Krisenkommunikation. Sprache in Zeiten der COVID-19-Pandemie. Baden-Baden.

Wengeler, M./Roth, K. S. (eds.) (2020): Corona. Essayistische Notizen zum Diskurs. (= Themenheft aptum. Zeitschrift für Sprachkritik und Sprachkultur 16, Issue 02/03). Hamburg.

Wiegand, H. E. (1989): Die lexikographische Definition im allgemeinen einsprachigen Wörterbuch. In: Hausmann, F. J./Reichmann, O./Wiegand, H. E./Zgusta, L. (eds.): Wörterbücher. Ein internationales Handbuch zur Lexikographie, Vol. 1. Berlin/New York, pp. 530–588.

Wiegand, H. E. (1998): Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie. Vol. 1. Berlin/New York, pp. 47–58.

Contact information

Annette Klosa-Kückelhaus
Leibniz-Institut für Deutsche Sprache
klosa@ids-mannheim.de

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Carolin Müller-Spitzer/Jan Oliver Rüdiger

THE INFLUENCE OF THE CORPUS ON THE REPRESENTATION OF GENDER STEREOTYPES IN THE DICTIONARY. A CASE STUDY OF CORPUS-BASED DICTIONARIES OF GERMAN

Abstract Dictionaries are often a reflection of their time; their respective (socio-)historical context influences how the meaning of certain lexical units is described. This also applies to descriptions of personal terms such as *man* or *woman*. Lexicographers have a special responsibility to comprehensively investigate current language use before describing it in the dictionary. Accordingly, contemporary academic dictionaries are usually corpus-based. However, it is important to acknowledge that language is always embedded in cultural contexts. Our case study investigates differences in the linguistic contexts of the use of man and woman, drawing from a range of language collections (in our case fiction books, popular magazines and newspapers). We explain how potential differences in corpus construction would therefore influence the "reality"[1] depicted in the dictionary. In doing so, we address the far-reaching consequences that the choice of corpus-linguistic basis for an empirical dictionary has on semantic descriptions in dictionary entries. Furthermore, we situate the case study within the context of gender-linguistic issues and discuss how lexicographic teams can engage with how dictionaries might perpetuate traditional role concepts when describing language use.

Keywords Gender linguistics; corpus-based lexicography; collocations; lexicography equality; gender equality

1. Have you ever googled 'woman'?

In 2019, the British PR manager Maria Beatrice Giovanardi wrote a blog post titled "Have you ever googled 'woman'?" in which she primarily complained about the description of women in various dictionaries, including lexicographic works by Oxford University Press, e. g. that *filly*, *biddy* or *bitch* are listed as synonyms for *woman*:

> The first search involved googling 'woman synonyms' and boom – an explosion of rampant sexism. I thought to myself, 'What would my young niece think of herself if she read this?' [...] Should data about how language is used control how women are defined? Or should we take a step back and, as humans, promote gender equality through the definitions of women that we choose to accept? [...] We talked about how the dictionary is the most basic foundation of language and how it influences conversations. Isn't it dangerous for women to maintain these definitions – wof women as irritants, sex objects and subordinates to men? (Giovanardi 2020)

She then started a petition at change.org, which was signed by 30,000 people. Oxford University Press responded by sending Katherine Connor Martin the following statement via The Guardian newspaper: The dictionary editors "are taking the points raised in the peti-

[1] What can be seen as "linguistic reality" is a very complex matter that goes beyond the scope of this paper. When we use in the following the term "linguistic reality", we are aware that texts or corpora are not a "description" or "representation" of this assumed reality, but serve to construct and interpret one possible part of this reality from language use (e. g. simply in reading or in specific work such as lexicography).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

tion very seriously […] As ever, our dictionaries strive to reflect, rather than dictate, language so any changes will be made on that basis. (Flood 2019). Here, reference is made to the descriptive tradition of modern lexicography. But in our view, two questions arise from this statement: a) What is regarded as a basis for the 'reflection of language'? In the tradition of modern corpus-based lexicography, it is the underlying corpus 'base'. But does everything from this corpus base always have to be included in the dictionary? Or should it rather be a curated selection? b) Should language use find its way into the dictionary, even if it could perpetuate gender stereotypes that, at least in part, no longer fit with contemporary ideas of society? Is it acceptable to reproduce racist and sexist attitudes exactly as they are (still) used?

2. "The man's a genius" and "she's really a nice woman": gender stereotypes in dictionaries

Dictionaries are often a reflection of their time, i. e. how they describe the meaning of certain lexical units must always be seen in their respective historical context. They are one of the sources to reflect gender roles (Nübling 2010, p. 594) for the first time, the lexicographic construction of gender in more recent editions of German dictionaries (from 1980 onwards. Consider the following example phrases taken from the entries on *man*, *woman*, *girl* and *boy* in the Cambridge Dictionary, reproducing stereotypical gender concepts:[2]

- "He plays baseball, drinks a lot of beer and generally acts like one of the boys."
- "Steve can solve anything – the man's a genius."
- "She's a really nice woman."
- "Who was that beautiful girl I saw you with last night?"
- "Both girls compete for their father's attention."

We understand stereotypes as thinking in group categories, although we acknowledge that this topic is treated in a much more differentiated way in social psychology:

> Indeed, individuals and groups can be said to be the central facts of society. Without individuals there could be no society, but unless individuals also perceive themselves to belong to groups, that is, to share characteristics, circumstances, values and beliefs with other people, then society would be without structure or order. These perceptions of groups are called stereotypes. (McGarty/Yzerbyt/Spears 2002, p. 1)

Such group descriptions concerning gender can be found in many dictionaries. Very pointedly and amusingly, Luise Pusch has shown this for example phrases of the German Duden dictionary of meanings from 1970:[3] The man, i. e. "he", "shows an acrobatic mastery of his body", "his soul is able to encompass the universe" and "great effect emanated from him". "She," on the other hand, "is always neatly dressed," "took the baby out daily," "awaits his return with great anxiety," and "she looked up to him as to a god."[4] Pusch summarizes: "In the preface, the editors write that the 'basic vocabulary of German in its basic meanings' is

[2] https://dictionary.cambridge.org/de/.

[3] Duden Bedeutungswörterbuch, Mannheim 1970.

[4] Original: Der Mann, also „er", „zeigt eine akrobatische Beherrschung seines Körpers", „seine Seele vermag das All zu umfassen" und „große Wirkung ging von ihm aus". „Sie" dagegen „ist immer adrett gekleidet", „hat das Baby täglich ausgefahren", „erwartet mit großer Angst seine Rückkehr" und „sie sah zu ihm auf wie zu einem Gott".

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

to be presented. They succeed in much more: they convey a deep, unforgettable insight into the soul of German, into its basic treasure of feelings and thoughts." (Pusch 1984, p. 144 [own translation]; cf. in more detail on various dictionaries of German Nübling 2010). This may illustrate that dictionaries are often a mirror of their time and thus also one of the important "platforms for productions of gender" (Nübling 2010, p. 594). Similarly, in their analysis of a contemporary Chinese dictionary, Hu/Xu/Hao (2019) point out that

> Women are often constructed in peripheral and domestic roles, as daughter, mother or grandmother. Their experiences are mostly restricted to themselves and their adjacent environment. When they act, their actions rarely bring noticeable changes to other participants or to the environment. Women are described as sensitive, loving and emotional, particularly preoccupied with familial, marital and domestic matters. On the other hand, men are mostly constructed in their central and social roles, as the prototypical adult men. […] Men are described as strong in physical strength, versatile in skills and noble in their actions. In other words, men are represented as valuable, active social members. (Hu/Xu/Hao 2019, p. 28)

Regardless of whether one sees this as an adequate description of 'reality' or as an overly stereotypical representation of men and women, the question arises whether such representations of gender in dictionaries are or can be intentional. For example, John Sinclair states in the preface to the 1987 Collins Cobuild English Language Dictionary that they "have abandoned the convention whereby *he* was held to refer to both men and women." This was done for various reasons, including the fact that "it is a very sensitive matter for those who have pointed out the built-in sexism of English" (Sinclair 1992, p. XX). This conscious positioning is particularly relevant for dictionaries because they can be understood as normative instances, even if they are primarily intended to be descriptive:

> This brings up the question of usage and authority. These concepts must support each other or no-one will respect either of them. If their close relationship breaks down, and authority is not backed up by usage, then no-one will respect it. […] Similarly, no-one will respect usage if it is merely an unedited record of what people say and write. […] Any successful record of a language such as a dictionary is itself a contribution to authority. (Sinclair 1992, compare also: Ripfel 1989, p. 204; Barnickel 1999, p. 171; Hidalgo Tenorio 2000, p. 225; Kotthoff/Nübling 2018, p. 180)

Against this background, lexicographers have a special responsibility. After Pusch's essay cited above, attempts were made in the Duden editorial office to improve the dictionary in many areas, e. g. to avoid unnecessarily stereotypical example phrases and to systematically include female occupational designations when they are common. (Kunkel-Razum 2004; for general comments, see Westveer/Sleeman/Aboh 2018). The main point here is to express awareness of the issue:

> Of course, dictionaries are not supposed to "straighten out" asymmetrical conditions that are solidified in the language system. It is undisputed and anchored in the German language (in the lexicon) that the entry *girl* always has to refer to the *easy girl* and the entry *boy* to the *tough boy*. It is not a matter of demanding a *heavy girl* or a *light boy* […]. Neither is it about *pregnant men* and *female machos*. It is about lexicographic doing gender. […] the question of which position on a scale from undoing gender via doing gender to hyper-ritualized gender the dictionaries take, in other words, which "degree of dramatization" they adopt – and whether they possibly engage in such dramatization themselves. (Nübling 2010, p. 595 [own translation])

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

The representation of gender in dictionaries thus seems to be caught between language use and lexicographic-moral responsibility. In our paper, in addition to discussing how much the lexicographer must or should intervene in the description of language use, we first investigate whether language use is indeed uniform at all. This question is particularly pertinent because we discovered strongly stereotypical statements about men and women in the entries of a modern corpus-based dictionary of German. It was newly compiled and therefore did not contain any old example phrases, e. g. examples inherited from earlier editions or other, older dictionaries. This finding was our starting point to examine the question of the data basis of 'language use' reflected in dictionaries.

3. Case study: influence of the corpus base on collocation sets for *Mann* (*man*) and *Frau* (*woman*)

3.1 Current lexicographic practice in German dictionaries

The starting point of our case study is the observation that even in modern corpus-based dictionaries of German, e. g. *elexiko*,[5] the descriptions of entries such as *Mann* or *Frau* are more influenced by stereotypes than we expected. *elexiko* is compiled from contemporary sources and does not contain old examples. This is why we thought we might find a more 'modern' representation of *Mann* or *Frau* in the dictionary. However, this is not the case.

In *elexiko*, collocation sets are listed for each head word. In the case of *Mann* and *Frau*, selecting the most frequent collocators leads to strongly different representations. It is particularly striking that for *Mann*, the agent role constitutes the second collocation set ("What does a *man* do?"), whereas for *Frau*, the patient role ("What happens to a *woman*?") is listed second – an imbalance that some researchers have already criticized as 'doing' gender (Hidalgo Tenorio 2000; Nübling 2010; Hu/Xu/Hao 2019) as how bias itself may organize human beings' experience by means of language in use. There exist well-known cultural stereotypes associated with the male and female conditions, and it is necessary to acknowledge the limitations to the application of many an impressionistic linguistic study on such issues. Taking this into account, the aim of this paper is to look at the way certain aspects of present-day English (a natural-gendered language). The fact that these collocation sets are presented in the dictionary in this way is due to the frequency of the groups, i. e. the patient role of women is much more prominent in the corpus base of *elexiko* than the agent role. For men, it is the other way round. Within the collocation sets for "what is discussed in connection with *man* or *woman*?", *man* collocates with: *car*, *erectile dysfunction*, *fire department*, *soccer*, *equality*, and *handball*. For *woman*, it is *age*, *occupation*, *breast cancer*, *emancipation*, *employment*, *birth*, *children*, *sex*, and *menopause*.

The *elexiko* team expresses critical awareness of these stereotypical representations. They point out that in the case of *woman*, reference is often made to their social roles in the family context (*single parent*, *divorced*, *unmarried*) or their general employment status (*unemployed*, *employed*), whereas in the case of *man*, such characterisations are absent. Adjectives such as *armed*, *masked*, *suspicious*, *hooded* only appear in the entry *man*, probably because the newspaper-heavy corpus contains many reports of violence and crime (Klosa/Storjohann 2011, p. 64).

[5] elexiko (2003 ff.), in: OWID – Online Wortschatz-Informationssystem Deutsch. Ed. by Leibniz-Institut für Deutsche Sprache, Mannheim, http://www.owid.de/wb/elexiko/start.html.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Further stereotypical representations can be found in the computer-generated collocation profiles ("Typische Verbindungen (computergeneriert)") within the Duden Online website.[6] Typical adjectives for *man* are *young*, *old*, *rich*, *strong*, *adult*, *powerful*, *armed* and *right*, whereas the typical ones for *woman* are *young*, *old*, *beautiful*, *tall*, *naked*, *pregnant*, *gracious* and *employed*. The corpora on which the two dictionaries (*elexiko*, Duden Online) are based – like the large linguistic corpora on German in general – are dominated by newspaper texts and such corpora have already been criticised as unbalanced in the context of lexicography (cf. Rundell/Atkins 2013, p. 1339). This is particularly relevant in the case of the computer-generated collocation profiles in Duden because they are obtained from the current newspaper-heavy Duden corpus. The extent to which this corpus base influences the representations must be carefully scrutinized. Contrastingly, the example phrases are more strongly informed by earlier editions of the dictionary and manual lexicographic analysis.

Linguistic practice is always embedded in a cultural context. "Language exists only in its use, and this is always culturally framed; at the same time, cultural facts, cultural habits, conceptualizations, and values are constructed and sedimented – indeed, archived – through language and in language" (Günthner/Linke 2006, p. 19, own translation). The empirical basis of lexicographic work transfers this linguistic-cultural context, and therefore a particular perspective on the world, into the dictionary. For example, men may indeed be more criminal than women, and women are also raped by men – however, it is open to debate whether exactly these aspects of 'reality' should be the main perspective of dictionaries.

The German collocation dictionary provides another solution for this issue: The entries for *Mann* and *Frau*, as with *elexiko* and Duden, are also designed to represent language use, but they are clearly displayed in a parallel structure (see Fig. 1). This approach requires more manual post-editing of the corpus data (which might have other, also negative, implications). According to a colleague who worked on the dictionary, this was a conscious decision.

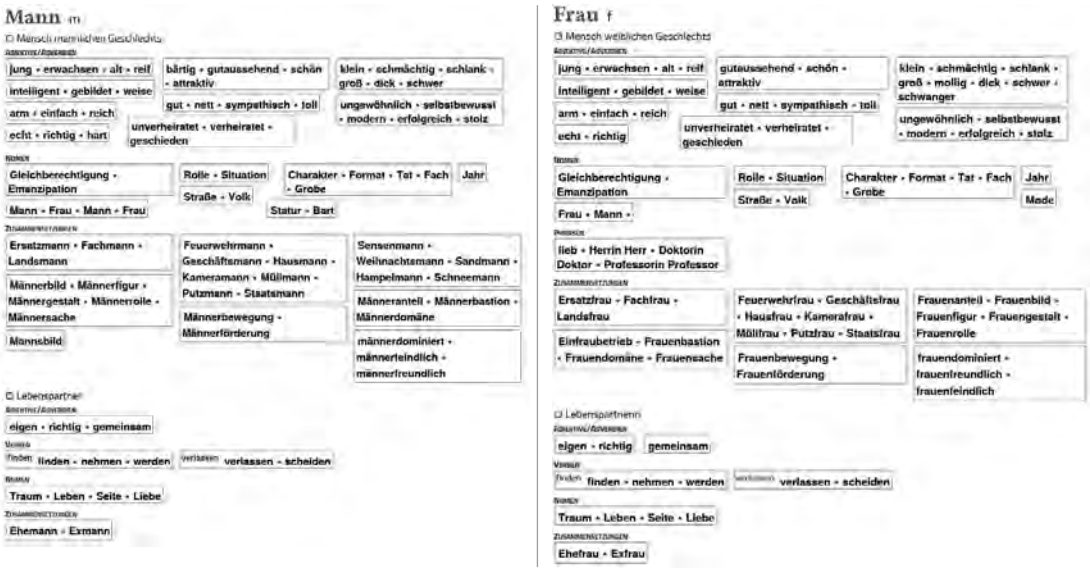


Fig. 1: Entries *Mann* (*man*) and *Frau* (*woman*) in the German collocation dictionary (Buhofer et al. (2015)

[6] www.duden.de.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

In a next step, we present a case study in which we investigate whether the collocation sets for *Mann* and *Frau* would change significantly if the corpus base was not predominantly composed of newspaper texts. We examine whether different corpus bases lead to different embeddings of *Mann* and *Frau*, addressing the urgent question of what we consider to "reflect language". We then discuss which methodological implications this could have for corpus-based lexicography in general. We end by addressing the fundamental question of how lexicographers should or could position themselves regarding the representation and perpetuation of gender stereotypes in dictionaries.

3.2 Method

The analyses presented in the following are based on three corpora constructed from different source materials:

- The corpus 'Fiction Books'[7] is based, as the name suggests, on various works of fiction (20th and 21st century). These corpora are listed in DeReKo[8] with the prefix 'LOZ-*'. Additionally, the corpora 'Mannheimer Korpus 1' and the 'THM – Thomas Mann Korpus' are included because they also consist of fictional texts.
- The '*elexiko*' corpus is based on the sources used for the '*elexiko*' dictionary (only newspaper texts), as well as more-recent newspaper-based documents (up to DeReKo Release 2021-I). Sources are: *St. Galler Tagblatt*, *Berliner Zeitung*, *Braunschweiger Zeitung*, *Burgenländische Volkszeitung*, *Bonner Zeitungskorpus*, *Deutsche Presse-Agentur*, *Tages-Anzeiger*, *Frankfurter Allgemeine*, *Handbuchkorpus*, *Hannoversche Allgemeine*, *Hamburger Morgenpost*, *Tiroler Tageszeitung*, *Kleine Zeitung*, *Berliner Morgenpost*, *Mannheimer Morgen*, *Salzburger Nachrichten*, *Niederöstereichische Nachrichten*, *Die Presse*, *Frankfurter Rundschau*, *Rhein-Zeitung*, *Der Spiegel*, *Die Südostschweiz*, *die tageszeitung*, *Vorarlberger Nachrichten*, *Oberösterreichische Nachrichten* and *Die Zeit*.
- The 'magazines' corpus consists of various periodical magazines. Sources are: *art*, *BEEF!*, *brand eins*, *BRIGITTE*, *Capital*, *Chefkoch*, *Couch*, *Eltern*, *Essen und Trinken*, *Gala*, *GEO*, *Living at Home*, *Nido*, *NEON*, *Psychologie Heute* and *Schöner Wohnen*.

	fiction books	**magazines**	**newspapers (*elexiko*)**
Time Range	1893–2011	2005–2020	1947–2020
Texts	1.320	60.066	15.831.499
Sentences	1.221.373	2.511.280	263.625.222
Tokens	22.132.897	37.771.792	4.398.207.319

Table 1: The three differently confected corpora for our case study

[7] One reviewer of the abstract correctly pointed out that we compare text types (newspapers, magazines) with a genre (fiction). Of course, fictional texts can also be found in newspaper texts, even if only to a small extent. However, we think that it is legitimate for our case study to proceed in this way, because corpus compilations in corpus linguistic practice usually include whole sources: whole newspapers, whole magazines or whole books. However, calling the fictional corpus a "book corpus" because we took fiction books seemed too general. By calling it "fiction books", however, we hope to have appropriately taken up the criticism.

[8] DeReKo: http://www.ids-mannheim.de/dereko.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

As can be seen in Table 1, the three corpora differ considerably, both in terms of the number of texts and the number of tokens. The three corpora also encompass different time periods. For the fiction books corpus, older texts were included a) because there are very few fiction books in the IDS corpora in general, and b) because limiting the corpus to recent texts would have resulted in too small a collection for the analyses required. The popular magazines such as Beef!, Brigitte Woman, Chefkoch or Living at Home are very recent, dating only from 2005 up to 2020. The *elexiko* corpus spans are wider time frame, namely from 1947 to 2020, but the largest amount of *elexiko* corpus texts can also be assigned to a very similar time period as the magazines. In the following, the *elexiko* corpus is referred to only as the newspaper corpus, since it consists exclusively of newspapers.

The corpora were imported into CorpusExplorer (Rüdiger 2021). For each search term (*Mann*/*Frau*), the corpora were separated so that only texts containing the particular search term were used for the co-occurrence calculation. A token-span (limit) for the calculation was not specified, and there was no restriction on parts of speech (POS). The sentence boundary was used to identify co-occurrences. Common co-occurrences were filtered by the 100 most significant entries (based on Poisson distribution, (cf. Heyer/Quasthoff/Wittig 2006, p. 134). To avoid visual distortion, we have filtered out the co-occurrences *young* and *old*, as their inclusion makes the observation and interpretation of the tag clouds more difficult.

3.3 Results

Figure 2 shows the most significant co-occurrences to *Mann* and *Frau* as a result of our analyses.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Fig. 2: All co-occurrences for *Mann* and *Frau* (the font size correlates with the significance based on Poisson distribution)

For the following comparison, the co-occurrences were filtered to include only tokens that were annotated at least once by the TreeTagger (cf. Schmid 1995) with the specific POS tag (here adjectives). In Figure 3 we see all adjectives co-occurring with *Mann* and *Frau* in our three corpora.

Fig. 3: Adjective co-occurrences for *Mann* and *Frau* (the font size depends on the significance based on Poisson-distribution)

Adjectives are used, among other things, to describe people. Thus, they would be included in a collocation set like "What is a *woman* or a *man* like?". What tendencies do our three corpora show in this regard?

Women are described in the fiction books partly regarding their external characteristics: *blond*, *pretty* or *attractive* (*blond*, *hübsch*, *attraktiv*), but also in terms of their marital status

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

(*married* or *divorced* – *verheiratet*, *geschieden*), or even *pregnant* (*schwanger*). The adjective *schweigestill* seems to us to be a tagging error (it is not a German adjective) and *gnädig* rather points to the quasi-lexicalized address *gracious woman* (*gnädige Frau*). In the newspaper texts, women are also described as *pregnant* (*schwanger*) or *working* (*berufstätig*), but also as being *raped* (*vergewaltigt*). Even in this case, however, passive constructions with participle 2 uses may be mixed with adjective uses. *Sexual* (*sexuell*) and *affected* (*betroffen*) could point to usage contexts such as *affected by sexual violence*. In the magazine texts, women are *self-confident*, *employed*, *attractive*, *pregnant*, *independent*, or *emancipated* (*selbstbewusst*, *berufstätig*, *attraktiv*, *schwanger*, *unabhängig*, *emanzipiert*). The magazines corpus is therefore the only dataset in which women are characterized by significant collocators that have nothing to do with their appearance or social role. We cannot efficiently classify the adjectives *feminine* (*weiblich*) and *masculine* (*männlich*); they are probably not used as direct attributes for women. The significant co-occurrence *sexual* is again likely to occur in bigger phrasal chunks, just like in the newspaper texts. What can be seen overall is that the collocation sets, as they would then be listed in the dictionary, would differ visibly depending on the corpus base.

The differences between corpora become even clearer with the adjectives co-occurring with *man*: in the fiction books, descriptive adjectives such as *gaunt*, *stout*, *stocky*, *bearded* or *lanky* (*hager*, *kräftig*, *untersetzt*, *bärtig*, *schmächtig*) dominate. *Dressed* (*gekleidet*) may not always be used as a direct attribute. In the newspaper texts, violent acts are a predominant topic. Logically, they are discussed more frequently in newspapers due to their news value: *armed*, *masked*, *alcoholized*, *previously convicted* (*bewaffnet*, *maskiert*, *alkoholisiert*, *vorbestraft*), but also more general words like *unemployed* or *powerful* (*arbeitslos*, *mächtig*). In magazines, men are described as *attractive*, *married*, *bearded*, *naked*, *gay* or **-looking* (*attraktiv*, *verheiratet*, *bärtig*, *nackt*, *schwul*, *aussehend*). Surprisingly, a considerable number of terms related to appearance, social role or sexual orientation are found here. The examples show clearly how differently 'linguistic reality' turns out, depending on which empirical basis is used.

Fig. 4: Verbal co-occurrences for *Mann* (the font size depends on the significance based on Poisson-distribution)

Similar differences appear in the verbal co-occurrences for *Mann*, i. e. fillers to collocation sets like "What does a *man* do?" or "What happens to a *man*?". Verbs in the fictional books are *muster*, *marry*, *observe*, *sit opposite*, *turn to* (*mustern*, *heiraten*, *beobachten*, *gegenübersitzen*, *zuwenden*; *erwachsen* is again a tagging error, *untersetzen* presumably also). In the newspaper texts, the context of violence is again predominant: *arrest*, *assault*, *threaten*, *shoot*, and *rape* (*festnehmen*, *überfallen*, *bedrohen*, *erschießen*, *vergewaltigen*) are particularly significant co-occurrences. In magazines, again, words referring to love life, money or power status are frequent collocators: *marry*, *fall in love*, *question*, *earn*, *cheat*, *question*, or *dominate* (*heiraten*, *verlieben*, *befragen*, *verdienen*, *betrügen*, *befragen*, *dominieren*). Again, the linguistic reality differs greatly.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

NN	newspapers	magazines
Frau		

Fig. 5: Nominal co-occurrences for *Frau* (the font size depends on the significance based on Poisson-distribution)

As a final example, we examine the nominal co-occurrences for *Frau*, i. e., "What is the topic of discussion in connection with *woman*?". For newspapers, the answer would be: *child*, *husband*, *violence*, *equality*, *social service* (*Kind*, *Ehemann*, *Gewalt*, *Gleichberechtigung*, *Sozialdienst*). For magazines, on the other hand: *leadership position*, *financial advisor*, *study*, *equality* (*Führungsposition*, *Finanzberaterin*, *Studie*, *Gleichberechtigung*; *percent* is more likely to be part of a phrase like "x percent of women are …."). Reflecting on language use would thus lead to very different results depending on the linguistic-thematic embedding of the words in the various text groups.

One should always keep in mind that co-occurrences say little about frequencies, but more about the strength of a connection. The fact that *woman* is so strongly associated with *gracious* (*gnädig*) in fiction does not mean that gracious women are often mentioned in total numbers, but that a (presumably low-frequency) word like *gracious* has a significant affinity to *woman*. Co-occurrences therefore indicate that certain activities or characteristics are strongly associated with women or men in the texts, which is more interesting for corpus-based research than mere frequencies.

3.4 Discussion and methodological implications

Our results show that in the newspaper texts, the common features of *women* and *men* as people who share many characteristics and actions step back in favour of the differences. The context of violence, for example, which is particularly over-represented in the *elexiko* entries,[9] is dominant only in the newspaper corpus. This is one of the instances where it becomes clear that the corpus basis can bring an unnecessarily strong bias towards *doing gender* into the dictionary (cf. also Nübling 2010, p. 620). This is especially problematic for lexicography:

> In fact, the question is to what extent a dictionary can involve a linguistic change; or, simply, whether its role in that process must be only one of perpetuation of what is actually supported by textual evidence; in other words, why a dictionary is allowed to repeat values which imply a biased representation of reality […]. (Hidalgo Tenorio 2000, p. 227)

Even if one assumes that a linguistic perspective always contains a "biased representation of reality", the case study has shown that the lexicographer chooses one of these linguistic views by selecting a specific corpus base, and that these linguistic perspectives on 'reality'

[9] In the entry *man* in *elexiko*, the first three verbal co-occurrences are *dominate*, *murder* and *shoot*.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

differ greatly. Gender stereotypes appear to be particularly strong in newspaper texts. These differences do not exist 'per se':

> There are not "the" gender differences in reality. […] This is neither to straighten out nor to idealize real relations nor to practice political correctness, but simply not to take a position on certain points – just as dictionaries do not take a position on racisms and anti-Semitisms (which can be found in reality as well as in corpora) by not reproducing them. (Nübling 2010, p. 628 [own translation])

In our opinion, it needs to be investigated more closely and discussed more intensively which implications go along with these findings. Our results show that different corpus bases lead to different linguistic representations of men and women, and that it should be best-practice to build dictionary entries on a diversified empirical base. However, more stratified compilation of the corpus may not be the best solution either, because it is then no longer possible to distinguish the different influences of the individual text groups. One possibility might be to at least refine the methods for analyzing vocabulary for a general dictionary, e.g. by performing co-occurrence analyses with different corpora containing different text types, and then comparing the resulting lists. This approach, according to our case study, is more likely to result in the most diverse representation possible. It would then also be possible to draw more precise conclusions about which texts have which influences. Our approach follows Sinclair's clarion callfor a very fine-grained documentation of all corpus data in order to be able to better interpret the results of corpus analyses:

> Also at any time a researcher may get strange results, counter-intuitive and conflicting with established descriptions. Neither of these factors proves that there is something wrong with the corpus, because corpora are full of surprises, but they do cast doubt on the interpretation of the findings, and one of the researcher's first moves on encountering unexpected results will be to check that there is not something in the corpus architecture or the selection of texts that might account for it. (Sinclair 2004, chap. 1)

Of course, this requires a very good lexicographic working environment, so that such procedures do not become too time-consuming. In any case, it becomes clear that the linguistic-technological methods cannot be used as a 'black box', but must be intellectually understood in order to be able to correctly classify the findings. The lexicographic work environment should make the variability of language use explorable.

4. Concluding remarks

Even though the orientation to actual language use in dictionary writing is certainly a very important principle of modern lexicography that has made dictionaries better tools, we believe that orientation to language use does not relieve lexicographers of their responsibility to take the political or social implications that language descriptions may have into account. As Nübling puts it: "Overall, one should assume that there is an awareness of gender constructions, especially in lexicographic teams, at the turn of the 20th and 21st centuries." (Nübling 2010, p. 609). A good compromise is certainly first to research language use with as much reflection (and self-reflection) as possible and then also – as one does with offensive or vulgar expressions – to find a compromise between language use orientation and the handing-down of outdated role models. We want to end with 'food for thought', citing David Foster Wallace's essay on "Authority and American Usage" in which he formulates the weak points of descriptive lexicography somewhat provocatively:

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

> But these flaws still seem awfully easy to find. Probably the biggest one is that the Descriptivists' "scientific lexicography" – under which, keep in mind, the ideal English dictionary is basically number-crunching: you somehow observe every linguistic act by every native/naturalized speaker of English and put the sum of all these acts between two covers and call it The Dictionary – involves an incredibly crude and outdated understanding of what *scientific* means. It requires a naïve belief in scientific Objectivity, for one thing. Even in the physical sciences, everything from quantum mechanics to Information Theory has shown that an act of observation is itself part of the phenomenon observed and is analytically inseparable from it. (Wallace 2001, p. 46)

References

Barnickel, K.-D. (1999): Political correctness in learners' dictionaries. In: Herbst, T./Popp, K. (eds.): The perfect learners' dictionary (?). Berlin/Boston, pp. 161–174. http://doi.org/10.1515/9783110947021.161.

Buhofer, A. H. et al. (2015): Feste Wortverbindungen des Deutschen. Kollokationenwörterbuch für den Alltag. In: Neuphilologische Mitteilungen 116 (1), pp. 242–244. https://www.jstor.org/stable/26372470.

Flood, A. (2019): Thousands demand Oxford dictionaries 'eliminate sexist definitions'. In: The Guardian, 17 September. https://www.theguardian.com/books/2019/sep/17/thousands-demand-oxford-dictionaries-eliminate-sexist-definitions (last access: 22-03-2022).

Giovanardi, M. B. (2020): Open letter calling on @OxUniPress to change their entry for the word "woman" #SexistDictionary. Change.org. https://www.change.org/p/change-oxford-dictionary-s-sexist-definition-of-woman/u/25841171 (last access: 17-03-2022).

Günthner, S./Linke, A. (2006): Linguistik und Kulturanalyse – Ansichten eines symbiotischen Verhältnisses/Linguistics and cultural analysis – aspects of a symbiotic relationship. In: Zeitschrift für Germanistische Linguistik (ZGL) 34 (1–2), pp. 1–27. http://doi.org/10.1515/ZGL.2006.002.

Heyer, G./Quasthoff, U./Wittig, T. (2006): Text mining: Wissensrohstoff Text: Konzepte, Algorithmen, Ergebnisse. (= IT lernen). Herdecke. http://deposit.dnb.de/cgi-bin/dokserv?id=2783785&prov=M&dok_var=1&dok_ext=htm.

Hidalgo Tenorio, E. (2000): Gender, sex and stereotyping in the Collins COBUILD English language dictionary. In: Australian Journal of Linguistics 20 (2), pp. 211–230. http://doi.org/10.1080/07268600020006076.

Hu, H./Xu, H./Hao, J. (2019): An SFL approach to gender ideology in the sentence examples in the Contemporary Chinese Dictionary. In: Lingua 220, pp. 17–30. http://doi.org/10.1016/j.lingua.2018.12.004.

Klosa, A./Storjohann, P. (2011):): Neue Überlegungen und Erfahrungen zu den lexikalischen Mitspielern. In: Klosa, A. (ed.): *elexiko*. Erfahrungsberichte aus der lexikografischen Praxis eines Internetwörterbuchs. (= Studien zur Deutschen Sprache 55). Tübingen, pp. 49–80. https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/5154.

Kotthoff, H./Nübling, D. (2018): Genderlinguistik: Eine Einführung in Sprache, Gespräch und Geschlecht (= Narr Studienbücher). Tübingen.

McGarty, C./Yzerbyt, V. Y./Spears, R. (2002): Social, cultural, and cognitive factors in stereotype formation. In: McGarty, C./Yzerbyt, V. Y./Spears, R. (eds.): Stereotypes as explanations: The formation of meaningful beliefs about social groups. New York, pp. 1–15. http://doi.org/10.1017/CBO9780511489877.002.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Nübling, D. (2010): Zur lexikografischen Inszenierung von Geschlecht. Ein Streifzug durch die Einträge von Frau und Mann in neueren Wörterbüchern. In: Zeitschrift für Germanistische Linguistik (ZGL) 37 (3), pp. 593–633. http://doi.org/10.1515/ZGL.2009.037.

Ripfel, M. (1989): Die normative Wirkung deskriptiver Wörterbücher. In: Hausmann, F. J. et al. (eds.): Wörterbücher – Dictionaries – Dictionnaires. Ein Internationales Handbuch zur Lexikographie. (= Handbücher zur Sprach- und Kommunikationswissenschaft 5.1). Berlin/New York, pp. 189–207.

Rüdiger, J. O. (2021): CorpusExplorer. Düsseldorf. http://corpusexplorer.de.

Rundell, M./Atkins, B. T. S. (2013): Criteria for the design of corpora for monolingual lexicography. In: Gouws, R. H. et al. (eds.): Supplementary volume dictionaries. An international encyclopedia of lexicography, pp. 1336–1343. http://doi.org/10.1515/9783110238136.1336.

Schmid, H. (1995): Improvements in Part-of-Speech Tagging with an Application to German. Proceedings of the ACL SIGDAT-Workshop. Dublin, Ireland.

Sinclair, J. (1992): Introduction. In: Collins cobuild English language dictionary. London, pp. XV–XXI.

Sinclair, J. (2004). "Developing linguistic corpora: a guide to good practice". https://users.ox.ac.uk/~martinw/dlc/chapter1.htm.

Wallace, D. F. (2001): Democracy, English, and the wars over usage. In: Harper's Magazine, pp. 39–58.

Contact Information

Carolin Müller-Spitzer
Leibniz-Institut für Deutsche Sprache Mannheim
mueller-spitzer@ids-mannheim.de

Jan Oliver Rüdiger
Leibniz-Institut für Deutsche Sprache Mannheim
ruediger@ids-mannheim.de

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Laura Pinnavaia

IDENTIFYING IDEOLOGICAL STRATEGIES IN THE MAKING OF MONOLINGUAL ENGLISH LANGUAGE LEARNER'S DICTIONARIES

Abstract The aim of this paper is to show how lexicographical choices reflect ideological thinking, singled out by Eagleton (2007) into the strategies of rationalizing, legitimating, action-orienting, unifying, naturalizing and universalizing. It will be carried out by examining two twenty-first century editions of each of the five English monolingual learner's dictionaries published by Cambridge, Collins, Longman, Macmillan, and Oxford. The synchronic and diachronic analyses of the dictionaries and their different editions at the macro-structural level (the wordlists) and at the micro-structural level (the definitional styles) will show how the reduction and change of data, derived from heterogeneous social and cultural contexts of language use, to abstract essential forms, involves decisions about the central and peripheral aspects of the lexicon and the meaning of words.

Keywords English monolingual learner's dictionaries; ideology; British twenty-first century lexicography

1. Introduction

There is no single and simple definition for Ideology. Generally, it refers to a material process of production of ideas, beliefs, values in social life. A less general meaning of ideology refers to ideas that symbolise life experiences of a socially significant class. In attending to the promotion and legitimation of the interests of a social group in the face of opposing interests, ideology also appears as a suasive device. Indeed, ideology can also signify ideas and beliefs which help to legitimate the interests of a ruling group or class by distortion and dissemination as naturalization and universalization. Whether positively or negatively connoted, ideology is above all a matter of discourse. It is especially the "relation between an utterance and its material conditions of possibility, when those conditions of possibility are viewed in the light of certain power-struggles central to the reproduction of a whole form of social life" (Eagleton 2007, p. 223).

Dictionaries are discourse. They tacitly shape our view of the structure of language. They give us insight into the power-struggles that are at the basis of social life. "Dictionaries represent a specific form of discourse embedded within broader discourses that represent knowledge of the world" (Benson 2001, p. 4). The knowledge of the world that the dictionary represents "is inscribed within the structured version of the language that the dictionary presents to the user. It is not simply a question of the content of the statements that the dictionary makes about the language. It is equally a question of the structures that make those statements possible" (ibid.).

> The knowledge of the world that dictionaries represent imply a set of structures that position one's own culture as a centre for the production and distribution of knowledge of other cultures, which are to various degrees peripheral to it" (ibid.). "The process of lexicographical representation, constrained by rules and principles of lexicographical practice, leads not to the production of a direct reflection of the language 'as it is', but to the production of a version of the language with

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

> a definite form and shape. This version of the language both represents and conditions our conceptions of what the language is, what it is made of and the ways in which its component parts are related to each other" (ibid, p. 8).

In Britain and USA no academies of the English language exist. Dictionaries stand in for such language academies. Dictionaries of English describe and prescribe language use, implicitly telling us what it is and what it is not, but they are ultimately the result of the decisions lexicographers must make at macro- and micro-structural levels. At the macro-structural level, they must decide upon the wordlist; at the micro-structural level, they must think about the definitional style, defining vocabulary, presentation of lexical and encyclopaedic information, choice of illustrative examples. In making such choices, lexicographers have thus a great responsibility towards their readership, because they frequently involve decisions about the central and peripheral aspects of the lexicon and the meaning of words, based on procedures that involve the reduction of data, derived from heterogeneous social and cultural contexts of language use, to abstract essential forms. If the ideological-discursive aspect of lexicography has been strongly argued with reference to English language dictionaries for native speakers (see Adams 2015, 2020; Benson 2001; Kachru/Kahanej 1995; Moon 1989), much less debate has ensued for English dictionaries for learners, who, unlike native speakers, are less inclined to their own opinion and/or language instinct and are in more need of ideological enlightenment.

It is therefore the aim of this essay to show if and how lexicographical choices in the compiling of wordlists and definitional styles in learner's dictionaries reflect ideological thinking singled out in Eagleton's terms (2007). In other words, we will analyse the makeup of five English monolingual learner's dictionaries by attempting to identify the following six strategies: the rationalizing strategy that provides plausible explanations for social behaviour which might otherwise be the object of criticism; the legitimating one that establishes one's interests as broadly acceptable; the action-oriented strategy that extends from an elaborated thought to the minutiae of everyday life; the unifying one that lends coherence to the group/classes holding it and bestows unity upon society; the universalizing strategy whereby values and interests that are specific to a certain time and place are projected as the values and interests of all humanity; and lastly, naturalization, whereby social reality is redefined by the ideology to become co-extensive with itself, in a way which occludes the truth that the reality in fact generated the ideology.

2. Method

To carry out this research, we focussed on the synchronic and diachronic examination of the macro-structural level (the wordlists) and the micro-structural level (the definitional styles) of the five English monolingual learner's dictionaries published by Cambridge (CALD), Collins (CCELD), Longman (LDOCE), Macmillan (MED), and Oxford (OALD). From a diachronic point of view, the latest editions of each dictionary were compared with an earlier edition of the same dictionary: CALD2 (2003) with CALD4 (2013); CCELD4 (2003) with CCELD9 (2018); LDOCE4 (2003) with LDOCE6 (2014), MED2 (2007) with MED (online), and OALD7 (2005) with OALD10 (2018). From a synchronic point of view, all the earlier editions were compared and so were all the latest ones. The temporal constraints that prevented the examination of all the words in the dictionaries led us to circumscribe the investigation to a select series of topics regarding the themes of daily life, business and jobs, clothing and fashion, computer technology, education, politics and government, religion, and society.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

As the dates of the publications show, the examination was restricted to the new millennium editions only. The reasons for this stem from methodological and socio-cultural concerns. From a methodological perspective, it seemed important to examine editions published roughly in the same period and that is within the first twenty years of the twenty-first century and with a 10- to 15-year gap between the latest and earlier edition. Had we decided to examine the latest editions with the first editions of each dictionary, there would have been a clear temporal unbalance between the two: the first editions of OALD, LDOCE, CCELD, CALD, and MED date back to 1948, 1978, 1987, 1995 and 2002 respectively. This explains further why we chose to compare the latest edition of MED with its second edition dated 2007: the comparison with the earlier edition would not only have implied a comparison between a first and last edition, but also a much wider temporal gap between editions compared to the other dictionaries.

Apart from the methodological inconsistency that this decision might have represented, choosing not to focus on first editions and/or pre-millennium dictionaries was also grounded upon the attempt to provide a similar social and cultural setting for the research, both lexicographically and ontologically speaking. Differences in the temporal gap between first and last editions would necessarily have determined marked differences between dictionaries in the composition of the wordlists due to the ontological changes in the use of English, with the more evident inclusion of neologisms and exclusion of obsoletisms in the dictionaries with a greater temporal gap between first and last editions. Differences in the temporal gap would also have determined major differences in the lexicographical method too. Learner's dictionaries have indeed come a long way since their first editions: lexicographers in the twenty-first century are much more aware than they were in the past of the importance of laying out entries clearly, defining them intelligibly and providing fitting examples. Having chosen to analyse twenty-first editions of the five dictionaries meant envisaging results that would depend on lexicographical choice rather than on lexicographical inexperience.

3. Results

The findings of the research will be presented in the following sections devoted to each ideological strategy.

3.1 Rationalizing

The first and foremost feature that emerges from the examination of the five diction-aries is how each categorizes or *rationalizes* its wordlist differently. Unlike MED and CCELD, CALD, LDOCE, and OALD classify their words rigorously into a vast array of topics. Indeed, it is possible to consult all three dictionaries by searching for words under topics that range from arts, business, clothes, colours, economics, education, food, games, geography, hard science, leisure, medicine, military, nature, philosophy, politics, religion, society, sport, to technology. That Cambridge, Longman and Oxford consider it important to classify lemmas according to their semantic domain is proven by the fact that topic searches are possible in the earlier editions as well as in the latest ones – albeit in different ways. Whilst the differences between the number and types of topics found in CALD2 and CALD4 are slight (apart from the introduction of biology, environment, government and politics in the last edition, the topics from the second edition to the last do not change whatsoever), the differences

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

between LDOCE4 and LDOCE6 and between OALD7 and OALD10 are far greater. Compared to the earlier editions, the latest ones not only introduce a more exhaustive list of topics, but also arrange them differently. If in OALD7, for instance, there is only one main topic entitled politics, in OALD10, politics is one out of seven subtopics (crime and punishment, law and justice, people in society, religion and festivals, social issues, wars and conflicts) that all belong to the main topic entitled politics and society. On the other hand, if folklore, mythology, occult, philosophy and religion all belong to the main topic of religion and thought in LDOCE4, in LDOCE6 folklore, mythology, occult, philosophy, religion, and religion and thought are all main topics. Compared to Collins and Macmillan that concentrate less in arranging their words into topics (in MED2 and CCELD9, it is possible to consult the dictionary by means of a limited subject area search only; MEDonline does not provide this search possibility), we might venture to say that, by allowing users to access the dictionary via topics, Cambridge, Longman, and Oxford provide an added perspective to their world of words.

It is a principle of descriptivist lexicography that dictionaries should not evaluate words by including some and excluding others. Whilst this is a main prerogative of historical dictionaries whose wordlists naturally continue to grow, the findings concerning our synchronic dictionaries have shown that the latest editions do not necessarily exclude words that appear in the earlier ones. Even in the short 10-15-year gap of analysis, the number of words in all the learner's dictionaries increases, as is clearly stated in their front matters: hundreds of new words have been added to CALD4 (p. viii); "a wealth of new words and meanings" to CCELD9 (p. xi); compared to OALD7, OALD10 has "added 2,000 words to the core list for advanced level students" (p. vi); and compared to LDOCE4, LDOCE6 (p. vii) "contains thousands more collocations and synonyms, as well as additional words and phrases". As far as MEDonline is concerned, given it has also been an open dictionary since 2009, "thousands of words and phrases have been added [,] about half of [which] have been "promoted" to become full entries in Macmillan Dictionary". Indeed, if in the mid-twentieth century lexicographers had to find compromises between the inclusion of new words and exclusion of old ones in view of the space restrictions that paper dictionaries imposed upon them (see Pinnavaia 2013), the unlimited space provided by the internet and by other electronic supports has eliminated the need to sacrifice words in twenty-first century dictionaries. In fact, the exclusion of dated words is no longer necessary. For example, items such as *bloomers* and *fatigues* that were out of fashion, or terms such as *cords*, *nylons*, and *tweeds* that were dated already by the beginning of the twenty-first century can still be found in all the latest editions. Thanks to technology, "how to decide what cannot be left out and how to compress that into the space available" (Landau 1991, p. 42) is a challenge that twenty-first century lexicographers should no longer have to face.

That said, there are words that certainly cannot be left out of synchronic dictionaries; namely, the ones that reflect the state of the language at one set moment in time. It explains why so many new words in the learner's dictionaries point to new scientific and technological achievements. However, the findings of this research have shown that the new words in the learner's dictionaries are not solely the record of an evolving ontological world: they also point to a new way of thinking. The inclusion of items of clothing such as *burqa*, *chador*, *hijab*, *salwar kameez* in the latest editions of the dictionaries suggest that there is a new concern on behalf of editorial teams to foster a change of attitude and acceptance of the other. Were it simply the reflection of a new ontological state, these words would not have existed in the earlier editions of the dictionaries, but they do in some and inconsistently:

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

burqa/burka, *chador* are included in LDOCE4, MED2, OALD7; *dirndl, hijab, salwar kameez* in MED2 and OALD7, none of them appears in CCELD4 and CALD2. The lack of consistency in the recording of these terms across the early editions shows that at the beginning of the century to include such terms was purely a lexicographical choice. The fact that they all appear in the latest editions points to a moral obligation lexicographers now have to acknowledge that English society is made up of people and customs having different geographical, cultural, and political origins. In the name of tolerance and acceptance, a wide spectrum of variety seems to be *prioritized* in these latest editions, which no longer just mirror "the prevailing cultural view of our society that science and technology are of the highest importance" (Landau 1984, p. 21).

3.2 Legitimating

Acknowledging that variation exists in society is not only conveyed in dictionaries by inclusion but also by descriptive labels. And even more than inclusion, descriptive labels legitimate evaluations. In representing a judgement on the items to which they are applied and on the categories to which they belong, labelled items indicate something that is peripheral to the norm, as opposed to unlabelled items that are the normative centre of language. The full range of information provided by labels available in English language dictionaries has been noted, among others, by Landau (1984), Quirk et al. (1985), Hausmann (1989), reiterated more recently, among other scholars, by Bergenholtz/Tarp (1995), Atkins/Rundell (2008), and Svensén (2009), and can comprise up to eleven types of restrictions covering the three macro socio–cultural functions of language: the ideational (etymological origins, temporal span, frequency of use, the region of use, subject field), the interpersonal (level of formality, if used by certain social groups only, the linguistic community's attitude, deviation from the cultural standard) and the textual (whether literary or poetic, if employed in written or spoken texts).

As the front matters report, all three types of labels are used in the five learner's dictionaries. As to their distribution, MED2[1] uses 44 labels, CCELD9[2] 33, OALD10[3] 31 labels, CALD4[4]

[1] MED2 has 12 style and attitude labels (formal, humorous, impolite, informal, literary, offensive, old-fashioned, showing approval, showing disapproval, spoken, very formal, very informal); 14 regional labels (American, mainly American, Australian, British, Canadian, Caribbean, East African, Indian, Irish, New Zealand, Scottish, South African, Welsh, West African), and 18 subject field labels. It is worth pointing out that we were not able to retrieve this information from MEDonline. We presume it remains the same however.

[2] CCELD9 has 21 labels of style and attitude (approval, dialect, disapproval, emphasis, feelings, formulae, formal, humorous, informal, literary, offensive, old-fashioned, politeness, rude, spoken, technical, trademark, vagueness, very offensive, very rude, written), 6 regional labels (American, Australian, British, Irish, Northern English, Scottish), and 6 subject field labels.

[3] OALD10 has 16 labels of style and attitude (approving, disapproving, figurative, formal, humorous, informal, ironic, literary, offensive, slang, specialist, taboo, dialect, old-fashioned, old use, saying) and 15 regional labels (Australian English, British English, Canadian English, East African English, English from Northern England, English from the United States, Indian English, Irish English, New Zealand English, North American English, Scottish English, South African English, South-East Asian English, Welsh English, West African English).

[4] CALD4 has 21 labels of style and attitude (abbreviation, approving, child's word/expression, disapproving, female, figurative, formal, humorous, informal, literary, male, not standard, offensive, old-fashioned, old use, polite word/expression, saying, slang, specialized, trademark, written abbrevi-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

30, and LDOCE6[5] 20. As far as the differences across editions are concerned, some dictionaries innovate more than others. CALD and LDOCE introduce slight changes: the attitudinal labels "approving" and "disapproving", inexistent in LDOCE4, are added to LDOCE6, whilst the temporal label "dated" in CALD2 is replaced by "old-fashioned" in CALD4, which also adds the regional labels "Indian English", "South African English" along with the label indicating the medium "written abbreviation". OALD introduces a few more changes: besides the replacement of "technical" with "specialist", OALD10 removes the examples of use to define the labels "offensive", and "taboo",[6] and changes the examples of use for "humorous", "slang", "old-fashioned".[7] The editions that differ most belong to CCELD, because CCELD9 provides new definitions and examples of use for the pragmatic labels "approval", "disapproval", "emphasis", "feelings", "formulae", "politeness", "vagueness".[8]

What is and what is not labelled is not always a clear reflection of the state of the language but what each editorial team considers important for its readership. That MED2 includes more subject field labels than other dictionaries (even though paradoxically access to the dictionary via such topics is limited in this edition and not possible in MEDonline) discloses the importance the editorial team places upon presenting specialist vocabulary, given that "4000 new items of specialist vocabulary" have been introduced in the English language in the last twenty years (MED2, p. viii). The intention to record the state of the English language in such fine detail is reflected also in the number of geographical labels that not only the MED but also the OALD includes, in order to provide a thorough "coverage of World English" (MED2, p. viii). Whilst these labels marking specialist and/or regional lexemes may be synonymous of a more descriptive lexicographical method (Verkuyl/Janssen/Jansen 2008), the labels pointing to the interpersonal and textual functions of language reflect a more prescriptive method which has characterised learner lexicography right from the outset.

Indeed, Rundell (1998, p. 337) reminds us how learner lexicography moves away from "the inappropriate model of the native-speaker's dictionary of 'record' towards a more 'utilitarian' lexicography in which the needs of the user take precedence over all other factors". The fact that all five dictionaries include an important number of labels regarding style and attitude points to a concerted action by all the editorial teams not just to describe the English language, but also to prescribe correct usage. This is particularly evident in OALD and CCELD that include more labels of attitude than the other dictionaries and make explicit their lexicographical plan in their front matters. In claiming that it has remained "true to the principles that Hornby established" (OALD10, p. vi), OALD legitimates the use of labels so "that the kind or style of English [learners] are using is right in that particular context"

ation), 8 regional labels (Australian English, Indian English, Irish English, Northern English, Scottish English, South African English, UK, US), and 1 subject field label).

5 LDOCE6 has 15 labels of style and attitude (approving, biblical, disapproving, formal, informal, humorous, literary, not polite, old-fashioned, old use, spoken, taboo, technical, trademark, written), 3 regional labels (American English, Australian English, British English) and 2 subject field labels.

6 In the definitions of the label "offensive" and "taboo" OALD7 respectively includes the examples *half-caste, slut* and *bloody, shit*.

7 For example, *humourous* in OALD7 is exemplified with the words *ankle-biter* and *lurgy*; in OALD10 with *fisticuffs* and *ignoramus*.

8 For example, the definition of *approval* in CCELD4 reads "you can choose words and expressions to show that you approve of the person or thing you are talking about, e. g. *angelic*"; in CCELD9 it reads: "the label *approval* indicates that you like or admire the person or thing you are talking about. An example of a word with this is *broad-minded*.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

(Hornby 1974, p. xxvi). Similarly, re-echoing Sinclair's concern that learners should be able to "distinguish between good and bad usage" (Sinclair 1987, p. xxi), CCELD announces that the ninth edition "will help [them] to understand not only the meaning of words but also how to use them properly in context" (CCELD9, p. xi). The differences between the number and types of labels across the five dictionaries shows that evaluating what to and what not to label is not always objective, but dependent upon the aims and scopes of each editorial team, which we have seen can oscillate between the descriptive and the prescriptive.

3.3 Action-orienting

In advising how to use language correctly, labels can go much farther than to prescribe. Labels of tone and register, which "explicitly indicate attitudes towards language use" (Stein 1997, p. 162), can indeed go as far as to proscribe linguistic behaviour, as the following, defined in the front matters of the five dictionaries, may clearly show:

(1) CALD2/4:

OFFENSIVE: very rude and very likely to offend people.

(2) CCELD4/9:

OFFENSIVE: likely to offend people, or to insult them; words labelled OFFENSIVE should usually therefore be avoided, e. g. cripple.

RUDE: used mainly to describe words which could be considered taboo by some people; words labelled RUDE should therefore usually be avoided, e. g. bloody.

VERY OFFENSIVE: highly likely to offend people or to insult them; words labelled VERY OFFENSIVE should be avoided, e. g. wog.

VERY RUDE: used mainly to describe words which most people consider taboo; words labelled VERY RUDE should be avoided, e. g. fuck.

(3) MED2/online:

IMPOLITE: not taboo but will certainly offend some people.

OFFENSIVE: extremely rude and likely to cause offence

(4) LDOCE4/6:

NOT POLITE: a word or phrase that is considered rude, and that might offend some people.

TABOO: a word that should not be used because it is very rude or offensive.

(5) OALD7/10:

OFFENSIVE: expressions that are used by some people to address or refer to people in a way that is very insulting, especially in connection with their race, religion, sex or disabilities; (e. g. half-caste, slut only in OALD7). You should not use these words.

TABOO: expressions that are likely to be thought by many people to be obscene or shocking. You should not use them. (e. g. bloody, shit only in OALD7).

The labels and definitions above show how each dictionary has a different *action-oriented* or proscriptive attitude towards bad language. While caution of usage is implicit in all the labels, even among the softest, such as NOT POLITE or IMPOLITE, there is no doubt that there is a stark difference between each editorial team's approach in the application of labels.

Starting with CALD, we can see it uses one label only, OFFENSIVE. With this label lexicographers signal an unequivocal unpleasantness of use, but no explicit prohibition is made. In all the other dictionaries a prohibition is instead more or less declared. MED is prohibitive

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

in a covert way: the prohibition emerges from the definition of the label OFFENSIVE that contrasts with IMPOLITE which is defined as "not taboo". The distinction between what is 'sayable' and 'not sayable' appears greater in LDOCE than in MED, because NOT POLITE signals a disagreeable word that could cause moral injury, unlike TABOO that explicitly prohibits usage.

More proscriptive still are the remaining two dictionaries. OALD uses two labels that are both powerful. Whilst in MED and LDOCE the two labels distinguish more offensive words from less offensive ones, in OALD OFFENSIVE is opposed to TABOO to indicate differing semantic areas of offence: the former signals words that offend in relation to sensitive issues; the latter words that are insulting because they are rude. In both cases users are warned not to use them. Like OALD, CCELD also pursues this typological distinction. In fact, it uses the label RUDE to highlight the words considered insulting because shocking, and the label OFFENSIVE to highlight those that are discriminatory. Although this semantic distinction is not as explicit here as it is in OALD, it is nonetheless made clear through the examples that support these definitions.

Unlike OALD, but like LDOCE and MED, CCELD also takes into account the degree of insult and offence. Words that are more than just offensive or rude are labelled as VERY OFFENSIVE or VERY RUDE. CCELD is the most precise of the five dictionaries in supplying labels. Not only does it distinguish different degrees of offense, as do MED and LDOCE, but it also takes into consideration the two typologies of offence that only OALD differentiates. In the definition of all four labels, CCELD lexicographers declare that words thus labelled should be avoided and emphasize it by means of a warning symbol.

The use of 'harsher' labels undoubtedly discloses a stronger action-oriented strategy by lexicographers, and there is no doubt that, of all the dictionaries examined, CCELD is the most proscriptive. Whilst this censorial attitude may seem to contrast with the principle of descriptive lexicography, it may also be interpreted as a conscious action of responsibility and protection in a highly judgmental society. Aware that bad language is "an area of usage where great skill and judgement are required for effective use", already in the first edition Sinclair (1987, p. xx) deemed it fundamental to warn his non-expert readers that "rude, offensive, obscene, or insulting words should be treated with great care" (ibid.). Sinclair's concern and voice evidently continues to inform CCELD's twenty-first century editors too.

3.4 Unifying

If, on the one hand, applying labels gives editorial teams voice that may even at times hark back to the founding fathers of learner lexicography, on the other hand, the way word senses are set out and examples of use provided seems to take it away from them. Because twenty-first century lexicography demands that "word senses and examples of use [be] abstractions from clusters of corpus citations" Kilgarriff (1999, p. 1), the lexicographer's role in the construction of each entry appears less incisive, as if it were the dictionary speaking and not the compiler. As shown below, twenty-first century learner's dictionaries list every word sense of a lemma.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

(6) CALD4[9]

1. SPEAK. To pronounce words or sounds to express a thought, opinion, or suggestion, or to state a fact or instruction
2. THINK. To think or believe
3. to give as an opinion or suggestion about something
4. to show what you think
5. when something or someone is said to be a particular thing, that is what people think or believe about them.
6. To give information in writing, numbers, or signs.

(7) CCELD4/9

1. When you say something, you speak words.
2. You use say in expressions such as I would just like to say to introduce what you are actually saying, or to indicate that you are expressing an opinion or admitting a fact. If you state that you can't say something or you wouldn't say something, you are indicating in a polite or indirect way that it is not the case.
3. You can mention the contents of a piece of writing by mentioning what it says or what someone says in it.
4. If you say something to yourself, you think it.
5. If you have a say in something, you have the right to give your opinion and influence decisions relating to it.
6. You indicate the information given by something such as a clock, dial, or map by mentioning what it says.
7. If something says something about a person, situation, or thing, it gives important in-formation about them.
8. If something says a lot for a person or thing, it shows that this person or thing is very good or has a lot of good qualities.

(8) LDOCE4/6:

1 EXPRESS SOMETHING IN WORDS to express an idea, feeling, thought etc. using words

2 GIVE INFORMATION to give information in the form of written words, numbers, or pictures – used about signs, clocks, letters, messages etc

3 MEAN [transitive] used to talk about what someone means

4 THINK THAT SOMETHING IS TRUE used to talk about something that people think is true

5 SHOW/BE A SIGN OF SOMETHING to show clearly that something is true about someone or something's character

6 SPEAK THE WORDS OF SOMETHING to speak the words that are written in a play, poem, or prayer

7 PRONOUNCE to pronounce a word or sound

8 SUGGEST/SUPPOSE SOMETHING used when suggesting or supposing that something might happen or be true

(9) MED2/online

1. express something using words
2. have opinion
3. mean something
4. give information/orders
5. show what someone/something is like

[9] CALD2 includes only four word senses (speak, think, give information, expression).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

6. imagine something happening
7. use something as example
8. tell someone to do something

(10) OALD10[10]

1. speak
   to speak or tell somebody something, using words
2. repeat words
   say something to repeat words, phrases, etc.
3. give written information
   (of something that is written or can be seen) to give particular information or instructions
4. express opinion
   to express an opinion on something
5. show thoughts/feeling

to make thoughts, feelings, etc. clear to somebody by using words, looks, movements, etc.

6. show what somebody/something is like
   [transitive] to show, sometimes indirectly, what somebody/something is like
7. give example
   [transitive, no passive] to suggest or give something as an example or a possibility

By giving space to each word sense of the verb *say*, all five reference works reflect a whole new method of dictionary-making that is based on the theory that every lexico-grammatical structure has a meaning (Firth 1957). Associative senses are no longer grouped under denotational senses, as was more common in the early days of learner lexicography and before the onset of corpus linguistics (Pinnavaia 2013). To do so now would probably be regarded as unnecessary tampering with the state of the language. Indeed, lexicographers are seemingly much less conspicuous than they used to be, generating wordlists based on frequency counts (Dohi 2001, p. 1), and providing definitions based on examples of use extrapolated from corpora. This has not only changed learner lexicography, but – as can be seen from the way the lemma *say* is treated – has attributed a more or less *unified* identity to all five dictionaries.

3.5 Naturalizing and universalizing

The *unified* identity that the twenty-first century learner's dictionaries take on, owing to the homogenous way in which word senses are dealt with, is further endorsed by the way illustrative material is handled. Just like word senses, examples of use are included in such a way as to seemingly free "the lexicographer from responsibility for the construction of the example" (Benson 2001, p. 96). The examples of use that illustrate the lemma *marriage* in all five dictionaries are a case in point:

(11) CALD2/4:

They had a long and happy marriage

She went to live abroad after the break-up of her marriage

(12) CCELD4/9:

In a good marriage, both partners work hard to solve any problems that arise.

His son by his second marriage lives in Paris.

[10] In OALD7, the word senses are the same but presented in a different order.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

(13) LDOCE4/6:
She has three daughters from a previous marriage.
In Denmark they have legalized marriage between gay couples.

(14) MED2/online:
A long and happy marriage
Too many marriages end in divorce.

(15) OALD7/10:
A happy/an unhappy marriage
All of her children's marriages ended in divorce.

As can be read above, the lexicographers of all five dictionaries illustrate marriage in a similar manner: a happy and long relationship that can also end in divorce. The examples offer a very generic picture of the positive and negative aspects of marriage. Unlike the more stereotypical examples one may find in twentieth-century learner's dictionaries (see Pinnavaia 2013, p. 24), these examples, moreover expressed as declarative sentences, not only sound neutral and objective but also authoritative (see Wenge 2016, p. 328). By distancing the lexicographer's voice, the dictionary, appears to be *natural and universal*: a spontaneous, inevitable, and unalterable instrument that reifies social life.

4. Conclusions

The natural, universal, and unified structures of the five learner's dictionaries, created by their similar layouts and the seemingly invisible nature of their compilers, underscores the well-founded claim that dictionaries hide ideologies. Indeed, just like dictionaries for native speakers, these dictionaries for non-native speakers "are surrounded by myths of 'objectivity'" (Benson 2001, pp. 4 f.) that not only lend them credibility but also and more importantly authority. Even though the corpus-based method of extracting linguistic information has made twenty-first century lexicography a more objective practice, there is no doubt that lexicographical teams are still diffusing ideas and social tendencies in different ways. In this essay, we have attempted to show that different choices made regarding the macro and the micro-structures of two twenty-first century editions of each of the five learner's dictionaries CALD, CCELD, LDOCE, MED, and OALD can be likened to the ideological strategies that Eagleton (2007) identifies as rationalizing, legitimating, action-orienting, unifying, naturalizing and universalizing.

Indeed, we have shown how the differing number, denomination, and arrangement of topics by which each dictionary divides up its wordlist, along with the different number and types of labels each applies to define words is a clear reflection of the policies and intents of each lexicographical team and the publishing house it belongs to. In fact, whilst the decision to access words via semantic topics can certainly have important advantages for the language learner, as in the case of CALD, LDOCE, and OALD, it cannot be considered a stick that measures the reliability or descriptiveness of a dictionary. It simply stems from a procedure that prioritizes and rationalizes the truth in the interest of the dictionary-user. Similarly, by restricting words to certain uses, labels almost always "represent the views and prejudices of the established, well–educated, upper classes", whose prerogative it is to preserve and make legitimate their own use of the English language (Landau 1984, p. 303). Moreover, the fact that each learner's dictionary decides what to and what not to label and accompanies each label with a personalised explanation of its meaning endorses the introspective nature

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

of this lexicographical classification, and reveals the position dictionaries take in the description of language. In this research, CALD and MED seem to pursue a more objective approach, while LDOCE, OALD, and CCELD an increasingly more subjective and action-oriented one.

Although there is need for more detailed research of the five learner's dictionaries in order to examine more editions and more topics, we would like to conclude by saying that even though the strategies of rationalizing, legitimating, action-orienting, unifying, universalizing, naturalization have been seen to involve the making of all five learner's dictionaries, the differences between the dictionaries are stark. Produced by different publishing houses with different editorial policies, the five dictionaries are indeed far from being homogenous and none of them displays the whole truth regarding the English language, despite the impression each one may give. Indeed, for non-native speakers any one of these dictionaries often becomes a central and determining point of reference for the reception and the production of the English language. Consequently, each lexicographical team has a great responsibility towards this readership who, unlike native speakers, is less able to disentangle objectivity from subjectivity. In meeting the specific needs of learners, lexicographical teams' actions should thus be deeply pondered and well planned, because, as is well-known, the more their choices are clear-cut, the more the ideas governing and the factors promoting them are heightened. In trying to balance the "dictates of [their] profession, the demands of the culture [they are] trying to portray, and of the people [they are] writing for" (Chabata/Mahvu 2005, p. 259), it is only natural that lexicographers disclose a world view of beliefs. It is important, however, that for non-native speakers whose English language instinct needs nurturing these beliefs be as impartial and as helpful as possible.

References

Adams, M. (2020): A fair road for stumps: language ideologies and the making of the dictionary of American English and the dictionary of Americanisms. In: Dictionaries: Journal of the Dictionary Society of North America 41 (2), pp. 25–59.

Adams, M. (2015): Language ideologies and the American heritage dictionary of the English language: evidence from motive, structure, and design. In: Dictionaries: Journal of the Dictionary Society of North America 36, pp. 17–46.

Atkins, B. T./Rundell, M. (2008): The Oxford guide to practical lexicography. Oxford.

Benson, P. (2001): Ethnocentrism and the English dictionary. London/New York.

Bergenholtz, H./Tarp, S. (eds.), (1995): Manual of specialised lexicography. The preparation of specialised dictionaries. Amsterdam.

CALD2 (2003): Cambridge advanced learner's dictionary. Cambridge.

CALD4 (2013): Cambridge advanced learner's dictionary. 4th edition. Cambridge.

Chabata, E./Mavhu, M. (2005): To call or not to call a spade a spade: the dilemma of treating 'offensive' terms in Duramazwi Guru reChiShona. In: Lexikos 15 (= AFRILEX-Reeks/Series 15), pp. 253–264.

CCELD4 (2003): Collins cobuild English language dictionary. Glasgow.

CCELD9 (2018): Collins cobuild English language dictionary. Glasgow.

Dohi, K. (2001): Dr. Thorndike's influence on learners' dictionaries. In: Dictionaries: Journal of the Dictionary Society of North America 22, pp. 153–162.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Eagleton, T. (2007): Ideology. London/New York.

Firth, J. R. (1957): Papers in linguistics 1934–1951. Oxford.

Hausmann, F. J. (1989): Die Markierung in einem allgemeinen einsprachigen Wörterbuch: eine Übersicht. In: Hausmann, F. J./Reichmann, O./Wiegand, H. E./Zgusta, L. (eds.): Wörtebücher /.../. Ein internationales Handbuch zur Lexicographie /.../. Erster Teilbd. Berlin/New York, pp. 649–657.

Hornby, A. S. (ed.) (1974): Oxford advanced learner's dictionary of current English. Oxford.

Kilgarriff, A. (1997): I don't believe in word senses. In: Computers and the Humanities 31 (2), pp. 91–113.

Kachru, B./Kahanej, H. (eds.) (1995): Cultures, ideologies, and the dictionary: sstudies in honor of Ladislav Zgusta. (= Lexicographica, Series Maior 64). Tübingen.

Landau, S. (1984): The art and craft of lexicography. Cambridge.

Landau, S. (1991): Approaches to meaning and their uses in lexicography. In: Dictionaries: Journal of the Dictionary Society of North America (13), 1991, pp. 91–114.

LDOCE4 (2003): Longman dictionary of contemporary English. Harlow.

LDOCE6 (2014): Longman dictionary of contemporary English. Harlow.

MED2 (2007): Macmillan English dictionary for advanced learners. Oxford.

MED (online): Macmillan English dictionary for advanced learners. Oxford.

Moon, R. (1989): Objective or objectionable? Ideological aspects of dictionaries. In: Knowles, M./Malmkjær, K. (eds.): Language and ideology. (= ELR Journal 3). Birmingham, pp. 59–94.

OALDCE7 (2005): Oxford advanced learner's dictionary of current English. Oxford.

OALDCE10 (2018): Oxford advanced learner's dictionary of current English. Oxford.

Pinnavaia, L. (2013): The changing language of monolingual dictionary discourse: a diachronic analysis of the Longman Dictionary of contemporary English corpus. In: Poppi, F/Cheng, W. (eds.): The three waves of globalization: winds of change in professional, institutional and academic genres. Newcastle-upon-Tyne, pp. 285–306.

Quirk, R./Greenbaum, S./Leech, G./Svartnik, J. (eds.) (1985): A comprehensive grammar of the English language. London.

Rundell, M. (1998): Recent trends in pedagogical lexicography. In: International Journal of Lexicography 11 (4), pp. 314–342.

Sinclair, J. (ed.) (1987): Collins cobuild English language dictionary. Birmingham.

Stein, G. (1997): Better words. Evaluating EFL dictionaries. Exeter.

Svensén, B. (2009): A handbook of lexicography: tthe theory and practice of dictionary-making. Cambridge.

Verkuyl, H./Janssen, M./Jansen, F. (2008): The codification of usage labels. In: van Sterkenberg, P. (ed.): A practical guide to lexicography. Amsterdam, pp. 297–311.

Wenge, C. (2016): The discoursal construction of the lexicographer's identity in a learner's dictionary: A systemic functional perspective. In: International Journal of Lexicography 30 (3), pp. 322–349.

Contact information

Laura Pinnavaia
University of Milano (Italy)
Laura.pinnavaia@unimi.it

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Petra Storjohann

THE PUBLIC AS LINGUISTIC AUTHORITY: WHY USERS TURN TO INTERNET FORUMS TO DIFFERENTIATE BETWEEN WORDS

Abstract This paper addresses the question of why we face unsatisfactory German dictionary entries when looking up and comparing two similar lexical terms that are loan words, new words, (near)-synonyms, or confusables. It explains how users are aware of existing reference works but still search or post on language forums, often after consulting a dictionary and experiencing a range of dictionary-based problems. Firstly, these dictionary-based difficulties will be scrutinised in more detail with respect to content, function, presentation, and the language of definitions. Entries documenting loan words and commonly confused pairs from different lexical reference resources serve as examples to show the shortcomings. Secondly, I will explain why learning about your target group involves studying discussion forums. Forums are a valuable source for detailed user studies, enabling the examination of different communicative needs, concrete linguistic questions, speakers' intuitions, and people's reactions to posts and comments. Thirdly, with the help of two examples I will describe how the study of chats and forums had a major impact on the development of a recently compiled German dictionary of confusables. Finally, that same problem-solving approach is applied to the idea of a future dictionary of neologisms and their synonyms.

Keywords Internet forums; synonyms; confusables; sense discrimination; problem-solving approach

1. Introduction

In any language, there are specific lexical terms which can cause confusion and uncertainties among native speakers and language learners. It is often loanwords, neologisms, synonyms, or paronyms (confusables) which can trigger doubts about their appropriate contextual use and their exact semantic differences because they have foreign origins (loanwords), new and unknown meanings (neologisms), semantic similarities (synonyms[1]), or commonalities in their lexical forms (paronyms) and because they can designate similar concepts. Whenever two words exist in a close semantic relationship or even in lexical competition with one another, they pose linguistic difficulties. In the past twenty years, language forums have established themselves as linguistic authorities which the public uses to judge instances of lexical uncertainty. Typically, a user posts a question with or without elucidating the contextual circumstances in which a lexical choice between two words is necessary. Typical replies include suggestions, intuitive responses, or copied dictionary entries, and these are further commented on by different users or referred to again by the initial user.

Examining online forums, we see an astonishing number of questions relating to language situations where someone is seeking advice on how to distinguish between two or more lexical items belonging to one of the aforementioned categories. Often, users consult forums after looking up words in a dictionary and experiencing various dictionary-based problems (Murphy 2013). These often concern insufficient information, lack of encyclopaedic knowledge, missing entries, specific emphasis on dominant meaning, or ignorance of language change. Hence, dictionaries are not always the most effective resources to solve problems of

[1] For a discussion on the notion of synonymy used in lexicography cf. Murphy (2013).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

language production. This even holds true for native speakers who are sufficiently competent to identify and reflect on the information given in entries (Chon 2008). Although German has a long-established lexicographic tradition of describing loan words, synonyms, and neologisms, there are only a few monolingual contrastive reference works, such as a paronym dictionary, which allow users to look up two lexical items simultaneously in order to compare their meanings and usage.[2] Evidently, there is a genuine need for contrastive dictionaries explaining semantic nuances, equivalent terms, and relatedness. Despite the fact that user studies have uncovered a number of insights into dictionary behaviour, skills, and consultation habits and that these studies have identified strategies of dictionary use in interactions with existing online dictionaries (e. g. cf. Müller-Spitzer 2014, cf. Lew 2015), little research has been carried out to investigate actual communicative needs and the linguistic queries associated with them, together with their corresponding answers. In fact, this is the only way to truly understand the potential target group for a linguistic resource, to identify their skill levels, and to develop innovative tools to ensure appropriate and reliable use of the resource in specific situations (cf. Storrer 2013).

In this paper, I will show why some well-known dictionaries fail to address common user queries. At the same time, I will show how we can overcome unsuccessful lexicographic habits by studying users' enquiries carefully. Finally, I will demonstrate how central conceptual ideas for an online dictionary of confusables ("Paronyme – Dynamisch im Kontrast") were derived from forums and effectively implemented during its planning phase and how they could be applied to the development of a future dictionary of German neologisms, synonyms, and loan words.

2. Dictionary-based problems and forums

Today, popular options among resources for language consultation include search engines, user-generated collaborative formats like Wiktionary, digitised and new online dictionaries produced by publishing houses (e. g. Duden), academic reference guides like DWDS, and NLP-based lexical tools (e. g. WortschatzLeipzig). Generally, users are accustomed to these but are not aware of the differences between them in terms of their underlying data, editorial processes, or their compilers' qualifications. Most online resources are characterised by typical dictionary-based problems, and users face a variety of challenges, e. g. the exhaustiveness and reliability of lexicographic details, the relationship between linguistic and extralinguistic information, the lack of (corpus) examples, how up-to-dated the data is, and the use of appropriate description style.[3] In the worst cases, dictionaries ultimately confuse users and cause vocabulary problems instead of solving them. Modes of presentation are rarely subject to criticism by dictionary users in chats, unless they prevent them from locating relevant pieces of information.

Users searching for synonyms, for example, do so for different reasons. Chon (2008) refers to these as "competence deficit word problems", which occur when a word or specific aspects of it are unknown. Searching for contextually appropriate lexical substitutes in dictionaries

[2] A new contrastive tool is WikiUnterschied.com, which compares wiktionary entries in a table format.

[3] We know from user studies of German online dictionaries by Müller-Spitzer (2014), contents and reliability are most crucial to users.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

or thesauruses is also a typical problem in situations of text production when native speakers and language learners are searching for lexical alternatives (Rundell 1999). In the context of English language learners and with respect to synonyms, Chon (2008, p. 24) points out that "successful language production depends considerably on the ability to make appropriate lexical choices in dictionary entries […]". Looking up synonyms is also essential in a situation of language reception when users are not familiar with a specific item, usually a loan word, technical term, or a new word. A typical query in forums might adopt the following style: *What is the difference between Grippe/Erkältung/Influenza* (*flu/cold/influenza*)? Besides learning about collocational and syntactic norms, getting a deeper understanding of differences means internalising semantic and encyclopaedic variations.

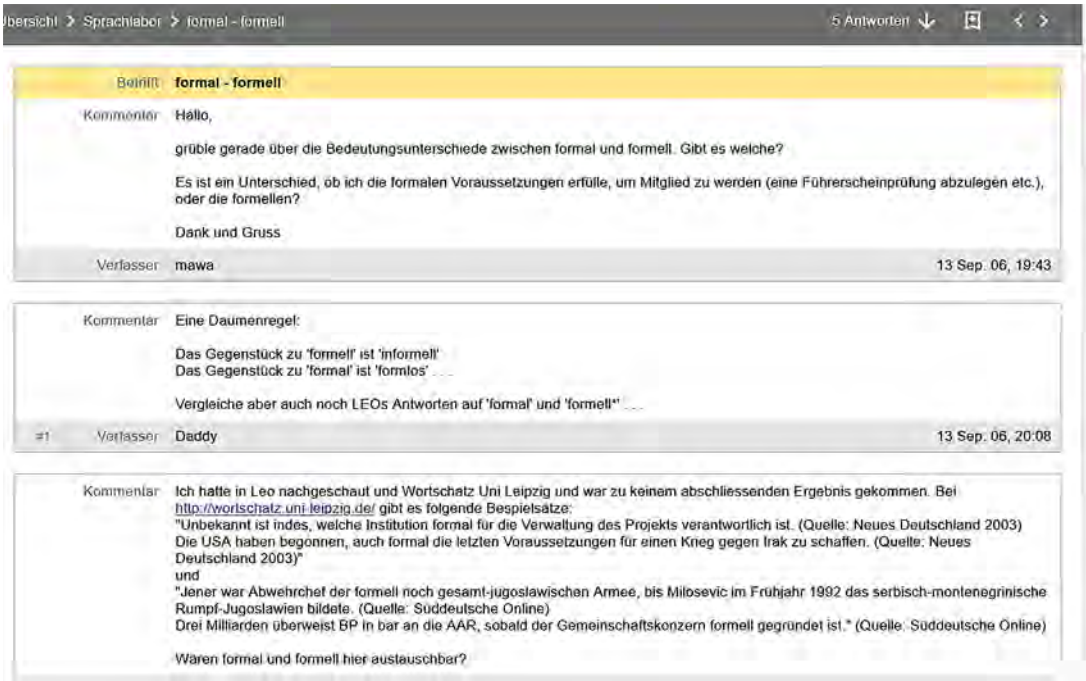
All these situations, at least to some degree, also apply to searching for easily misused words. Paronyms are similar to one another in their lexical form and often, to some extent, in meaning. They share a morphological root and typically differ with respect to prefixes or suffixes. A large number of paronyms are in fact loan words, such as *anarchisch/anarchistisch* (*anarchic*), *fiktiv/fictional* (*fictitious/fictional*), and some of them denote identical concepts and exist in well-established synonym relationships (e. g. *patriarchalisch/patriarchal/patriarchisch*, (*patriarchal*)). Using loan words, in particular, can cause misunderstandings, as they are stylistically marked and exhibit a certain degree of education. There are also terms with indigenous roots such as *farbig/farblich* (*coloured*, *colourful*, *in/concerning colour*) or *lesbar/leserlich* (*readable/legible*) which can cause problems. Again, these competence-deficit word problems often relate to both insufficient semantic and extra-linguistic knowledge. Speakers have different or only vague and subjective intuitions and show a lack of knowledge as to the precise contextual circumstances in which the terms should be used. In forums, questions like *What does autoritativ* (*authoritative*) *mean and how does is differ from autoritär* (*authoritarian*)? or *Is there is difference between fremdsprachig/fremdsprachlich* (*in terms of a foreign language?*) are a source of debate and controversy. With new words (coinings or new loan words) uncertainties differ. The element of novelty uncovers deficits in specific knowledge about a phenomenon. *What do the terms Covid/Corona/SarsCov-2 mean exactly?* is a question arising from new and simultaneous information and lexical input about similar or related phenomena.

In what follows (2.1 and 2.2), I will pick out common failings and pitfalls typically encountered when searching for lexical pairs with an explicit need to identify of a precisely drawn spectrum of meaning. I will look at their treatment in popular German dictionaries with regard to three aspects: lexicographic information, degree of detail, and defining style.

2.1 Depth and presentation of lexicographic information

In German, the use of *formal*, *formell*, or *förmlich* poses difficulties in various contexts. These loan words, adopted in the late 15th century from Latin *formalis*, are also paronyms and used synonymously in contemporary German in some of their contexts. Looking them up in a German dictionary is a confusing experience. In an example taken from the Leo-language forum (Fig. 1), a dictionary-based problem is reported by a user.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.


Übersicht > Sprachlaber > formal - formell 5 Antworten

Betrifft: **formal - formell**

Kommentar: Hallo,

grüble gerade über die Bedeutungsunterschiede zwischen formal und formell. Gibt es welche?

Es ist ein Unterschied, ob ich die formalen Voraussetzungen erfülle, um Mitglied zu werden (eine Führerscheinprüfung abzulegen etc.), oder die formellen?

Dank und Gruss

Verfasser: mawa 13 Sep. 06, 19:43

Kommentar: Eine Daumenregel:

Das Gegenstück zu 'formell' ist 'informell'
Das Gegenstück zu 'formal' ist 'formlos' ...

Vergleiche aber auch noch LEOs Antworten auf 'formal' und 'formell' ...

#1 Verfasser: Daddy 13 Sep. 06, 20:08

Kommentar: Ich hatte in Leo nachgeschaut und Wortschatz Uni Leipzig und war zu keinem abschliessenden Ergebnis gekommen. Bei http://wortschatz.uni-leipzig.de/ gibt es folgende Beispielsätze:
"Unbekannt ist indes, welche Institution formal für die Verwaltung des Projekts verantwortlich ist. (Quelle: Neues Deutschland 2003)
Die USA haben begonnen, auch formal die letzten Voraussetzungen für einen Krieg gegen Irak zu schaffen. (Quelle: Neues Deutschland 2003)"
und
"Jener war Abwehrchef der formell noch gesamt-jugoslawischen Armee, bis Milosevic im Frühjahr 1992 das serbisch-montenegrinische Rumpf-Jugoslawien bildete. (Quelle: Süddeutsche Online)
Drei Milliarden überweist BP in bar an die AAR, sobald der Gemeinschaftskonzern formell gegründet ist." (Quelle: Süddeutsche Online)

Wären formal und formell hier austauschbar?

Fig. 1: Question about the difference between *formal-formell*

His or her question referring to the distinction between *formal* and *formell* is put into a specific linguistic context where someone needs to fulfil official requirements in order to become a member or, for instance, in order to obtain a permit. In a second note the user adds "I checked Leo and Wortschatz Uni Leipzig and couldn't come to a conclusive result".[4] Then some examples are copied from the second resource and a further question follows "Can one mutually substitute *formal* with *formell*?". Figure 2 shows both entries in WortschatzLeipzig. Speakers will not successfully resolve their problems by using either entry in the NLP-tool, as they do not encounter a definition or any semantic information that can be used without further linguistic interpretation. Both items are polysemous and exhibit a range of semantic commonalities and differences. The entries, however, neither include senses and their differences nor correlate with any information about distinct usages. The examples lose their illustrative value when given as a block for a headword with many different options for contextual usage. Users cannot relate their existing knowledge and their specific query to this kind of entry without prior fine-grained disambiguation. The problem of assigning words to context is further increased by cross-referencing the headwords as synonyms and by referring to identical meaning equivalents (cf. Chon 2009, p. 28).

[4] Wortschatz Uni Leipzig is officially known as WortschatzLeipzig.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

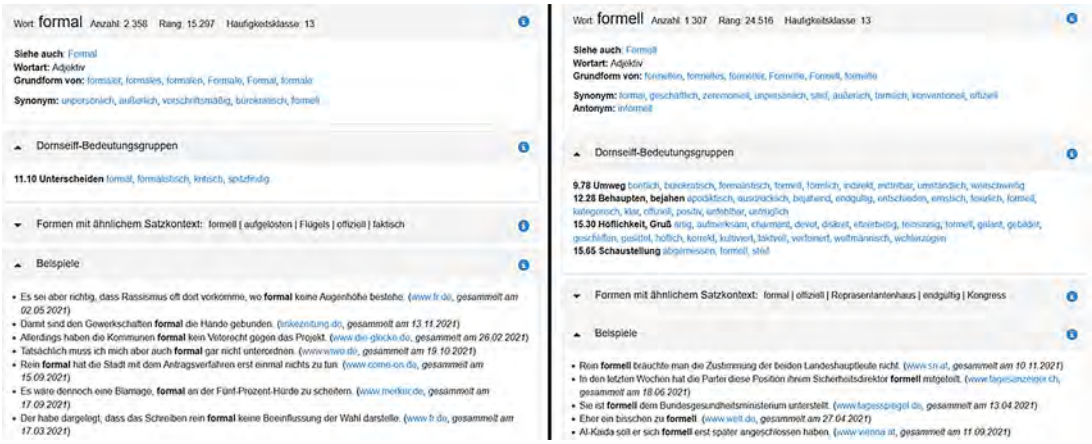


Fig. 2: Entries *formal* and *formell* in WortschatzLeipzig

Broadly speaking, the resource gives the impression that both items are almost identical, apart from the obscure fact that *formell* is embedded contextually in more thematic domains, as this lexeme is listed in four different meaning sets taken from the onomasiological dictionary Dornseiff whereas *formal* is only documented in one thematic group.

WortschatzLeipzig is a computer-generated tool and it appears to be used in chats as a source for common language queries. Identifying particular lexical environments and domains is a prerequisite to decide whether two terms are contextually interchangeable. As information is not adequately differentiated and presented and is not entirely reliable without underlying editorial procedures, the resource must be deemed unsuitable to answer the initial question.

2.2 Language of description and examples

One main criticism addressed in discussion forums is the language used in reference guides. Although user-friendliness and usability in terms of descriptive style has long been addressed in meta-lexicography, we still face some old problems.[5] Three difficulties can be observed. Firstly, abbreviations serving as usage notes to indicate register or variation are not always familiar to everyone or are difficult to decode. Secondly, we need to question the comprehensibility of a concise, logical, and structuralist style of definition which follows a strict formula of *genus proximum* and *differentiae specificae*. Such definitions do not correspond to everyday language use and were originally established for print dictionaries. They can cause a situation where looking up one term requires an endless series of additional terms to be looked up (cf. Antor 1994, pp 78 f.). This style of entry has often been adopted as an inherent dictionary style even for online resources where the lack of space is irrelevant. Alternatively, some dictionaries use single synonyms to paraphrase the term without further explanations of syntagmatic restrictions. This tradition goes back to the notion of the referential substitutability of words. Cobuild's dictionary for language learners was the first project with an informal and discursive style of definition that used basic vocabulary (cf. Hanks 1987). Only a few dictionaries (e.g. *elexiko*) have reflected carefully on adopting a different definitional style (cf. Storjohann 2005), avoiding abbreviations altogether within the German context. Thirdly, most definitions lack extra-linguistic information, which is

[5] For an overview see Rothe (2001).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

not a discrete category from linguistic knowledge about a word. As a result, a word's meaning cannot be fully explored when no allowance is made for its designated discourse properties or referential domains.

Some traditional definitions found today appear antiquated and awkward or are difficult to understand, so that the description provided fails to be received meaningfully by the user. As a result, users request explanations of meanings in "your own words" or concrete language examples to show words in context (see Fig. 3).

Fig. 3: Request for clarification of the difference between *legitim* and *legal* "in your own words"

The first request implies a discernible difference between *legitim* (*legitimate*) and *legal* (*legal*). The second request involves the search for examples to clarify an assumed difference between the two items. A closer look at the definitions and examples of the headwords in the DWDS, as summarised in Table 1, show both words are defined by synonyms which themselves contain the paronyms *gesetzmäßig* and *gesetzlich*.

legitim	***legal***
gesetzmäßig, rechtmäßig Beispiele: *eine legitime Macht, Regierung* *mit legitimen Mitteln arbeiten* [*jemand*] *der keinerlei Ansprüche stellte auf legitime Zuzugsgenehmigung* [*Kasack, Stadt, 579*]	dem Gesetz entsprechend, gesetzlich Beispiele: *eine legale Regierung, Partei* *etw. auf legalem Wege tun* *legal handeln* *Devisen legal erwerben, umtauschen*
ehelich Beispiele: *ein legitimer Nachkomme* *ein Kind für legitim erklären*	
Synonyms berechtigt · dem Recht entsprechend · erlaubt · gesetzeskonform · gesetzlich · legal · legitim · nach Recht und Gesetz · nach dem Gesetz · rechtens · rechtlich einwandfrei · rechtmäßig · statthaft · zugelassen · zulässig ● rechtssicher	**Synonyms** berechtigt · dem Recht entsprechend · erlaubt · gesetzeskonform · gesetzlich · legal · legitim · nach Recht und Gesetz · nach dem Gesetz · rechtens · rechtlich einwandfrei · rechtmäßig · statthaft · zugelassen · zulässig ● rechtssicher

Table 1: Dictionary information for *legitim* and *legal* in DWDS

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

As such, the user now actually needs to know the precise difference between two confusables. Synonyms paraphrasing a headword can be useful, but they need to be chosen carefully or further substantiated with additional lexicographic data. The example, e.g. *eine legitime Macht/Regierung vs. eine legale Regierung/Partei*, are quite similar, and the identical synonym groups at the end suggest a meaning overlap in at least one "shared" sense. The only difference recognisable is one additional sense ('ehelich' ('in wedlock')) for *legitim* when referring to humans. As will be shown in 4.2, *legitim* and *legal* are, in fact, not meaning equivalents at all. The definitional style and the examples used in the DWDS create an inadequate impression about their use.

Although the examples given in 2.1 and 2.2 refer to forums where native speakers exchange their thoughts, comparable questions are found in forums designed for language learners who address the difficulties they encounter when faced with learner dictionaries. Members of the general public participating in discussion forums recommend specific dictionaries and explain why they should be used or avoided. The answers also provide insight into speakers' intuitions, their linguistic and encyclopaedic knowledge, and their beliefs as well as their reactions to vague or strictly prescriptive suggestions. The best chats reveal the final decision on the lexical choice (and the reasons for it) based on different comments left in the forum.

3. Impetus for a new paronym dictionary

A few years ago, the Leibniz-Institut für Deutsche Sprache initiated a dictionary of confusables, the first corpus-assisted online guide to German paronyms. As far as German lexicography is concerned, it was the first time a dictionary project had based its lexicographic contents, design, and functionality on users' interests and expectations as derived from forums and by examining reports on individual instances of dictionary consultation (Storjohann 2016).[6] In the planning process, the project was interested in the target users, their linguistic competence, expectations, and experience with lexicographic data, and any conflicts with their own intuition etc. Through more cognitive-oriented studies of users we were able to include in the dictionary what users specifically demanded in their chats. Over 200 discussions on paronyms, including questions and reactions, were subject to examination. Specifically, our interest focussed on who showed uncertainties in their use of confusables, what the communicative contexts were in which difficulties occurred, and where users looked the words up. Once we learned about general dictionary skills, we analysed how satisfied the users were with the information in traditional entries and whether they differed from their own introspections. In addition, particular attention was paid to what skills are used to draw upon different types of knowledge and how users expressed a wish for more encyclopaedic information. Another fundamental question raised in the project was how users react to both vague and prescriptive answers and what choices they make when they receive a number of divergent responses.

[6] The results of this study were only used for design purposes during the development of the new resource. The project still holds all data (chats written between 2002 and 2016) from this investigation.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

The insights obtained played a central part in the planning process and led to new ideas and alternative lexicographic principles. One of the aims was to create a reliable and user-friendly tool by applying contrastive corpus-linguistic methods and by realising the demands of cognitive lexicography (e. g. Ostermann 2015). Another objective was to overcome some of the major dictionary-based problems by integrating innovative modes of presentation and by exploiting new technological possibilities. Sections 3.1 to 3.2 will provide a link to the challenges explained in 2.1 to 2.2 and show some lexicographic solutions to the lexical pairs mentioned in the forums. These mainly concern: how to quickly identify similarities and differences, how to combine sufficient linguistic and extra-linguistic knowledge, how to use new means of presentation, how to involve the user with interactive, adaptive functionality, how to choose a more accessible definitional style, and how to select examples best suited to illustrate context (and synonymy).

3.1 Depth and presentation of lexicographic information

The objective of producing a reliable source implies addressing contextual information in terms of ontological reference, collocability, and thematic domains in different contexts.[7] Overlaps and differences need to be clearly accessible and understood at first sight. Besides quickly accessing information, some users require further information which needs to be selective, customisable, and generated on demand. These prerequisites were put to the test a number of times in the initial stages of the dictionary. As a result, we created a two-level entry consisting of a contrastive overview and a more detailed level. Both levels contain interwoven lexical, semantic, and world knowledge about words, their senses, and conventions. Senses are understood to segment the overall meaning potential into meaningful units perceived as typical pattern choices from corpus analysis. In addition to developing ideas about contexts and depth of information and modifying the style of description, it was also essential to assess the technological options for presentation as well as the (visual) functionalities which assist in the design of the resource. In fact, forms of presentation and intelligent modes or functions allow for an efficient and intuitive navigational structure. They also support the explanations of the headword in many different ways, for example, by providing interactive guidance and user-adaptive choices and by changing the linguistic perspective.

The focal point of the contrastive overview are the headwords and their contextual uses (each in a tile) encompassing the full semantic spectrum of the word and signalling its context-boundedness to users as detected in the underlying corpus. The slots/positions and colour marking of the tiles help to identify the relationships between the senses of the corresponding partner term(s) (cf. Fig. 4).

[7] For a detailed account of the German paronym dictionary, see Storjohann (2018).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

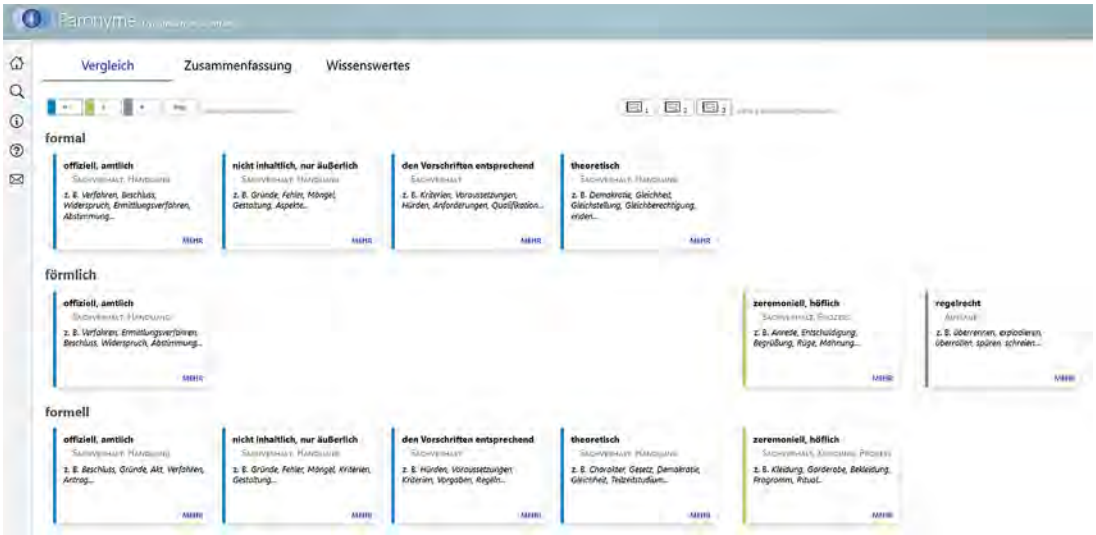


Fig. 4: Overview of senses of *formal*/*förmlich*/*formell* in the paronym dictionary

Formal/*förmlich*/*formell* are presented beneath each other with the most frequent term at the top of the entry. Their senses are in line horizontally and placed directly beneath each other when semantically identical or similar, but are offset when different. In cases where no equivalent sense exists, the allocated slots remain empty. The colour scheme further indicates whether senses are classified as being identical (blue), similar with slight semantic nuances (green), or distinct (grey). As such, the type of senses can be identified, arranged, and set into relationships with others. The semantic spectrum of the items is concisely depicted, and one can instantly grasp that the adjectives are polysemous with contextual overlaps and differences between them. A short description is provided for each sense, and the tiles also serve as points of contextual navigation to further detail. Looking at a tile more closely, it reveals the following information:

formal

(1) 'offiziell, amtlich'
(2) Sachverhalt, Handlung
(3) z. B. *Verfahren*, *Beschluss*, *Widerspruch*, *Ermittlungsverfahren*, *Abstimmung*

A synonym (or two) serves as a label for an identified context, while general reference is provided by ontological categories (e. g. STATE OF AFFAIRS, PERSON, PROCESS, SITUATION), and these are exemplified by collocates illustrating lexical realisations of the more abstract reference. Its content can be read as follows: *Formal* means 'official' (1) and it refers to STATE OF AFFAIRS and PROCESSES (2) such as a *procedure*, *decision*, *objection*, *investigation*, or *voting* (3). Together these discriminate sufficiently the contextual uses from each other. Via a menu, the senses can be rearranged flexibly depending on whether the focus is on commonalities, differences, or frequency. Knowing the precise circumstances under which both expressions (better: their senses) are interchangeable can require more detail. This can then be selected individually at the detail level, where information is more extensive and where users can study that detail side by side (Fig. 5).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

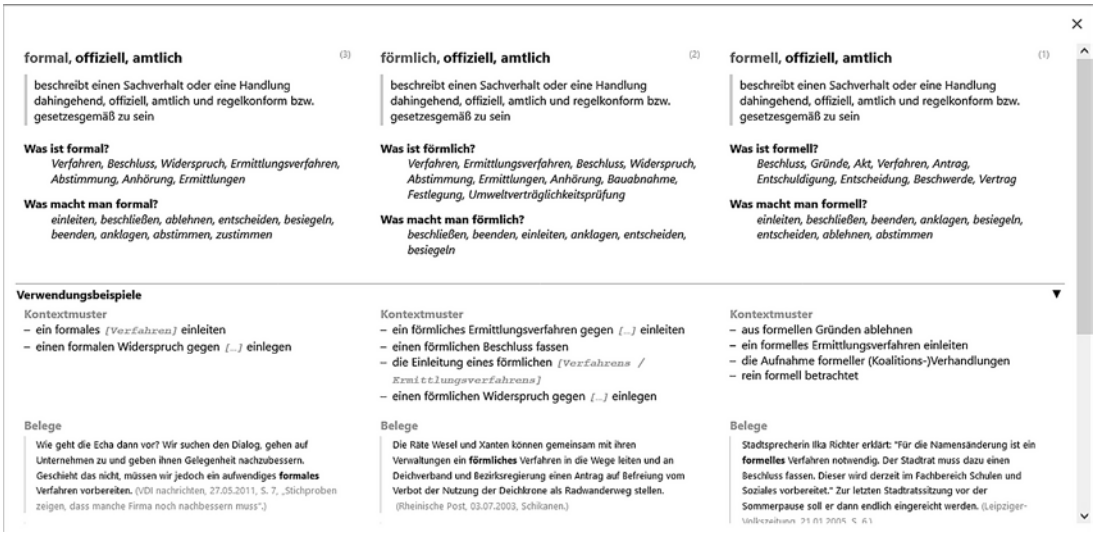


Fig. 5: Details of the sense 'official' shared by *formal/förmlich/formell*

The paraphrase here is longer, with the reference categories embedded into further relevant contextual information. More collocates open further contextual options, and these are classified according to word class to show their syntagmatic role (similar to semantic frame organisation). These also help the user to avoid violating conventional collocational patterns. Together, they create an interplay of lexical and non-lexical information. Corpus examples, typical construction patterns, and synonyms/antonyms allow for further comparison and illustration. As looking up paronyms often occurs in situations of text production, locating diverse and comprehensive information on a specific word is essential. For such activities, Lew (2015, p. 9) remarks:

> The lexicographic treatment should be more detailed than for text reception, allowing the dictionary user to construct natural phrases and sentences with the headword. To that end, the user will typically need guidance on syntactic patterns into which the headword enters, as well as collocates, preferably with examples of use to serve as a model for production.

Deciding what the essential type and the necessary depth of detail are, as well as where to present information and how to integrate sections generated on demand, has turned out to be highly complex also with respect to editorial practice. The editorial process includes the analysis and interpretation of corpus data, the discrimination of senses, the allocation of data to each sense, and the assigning of uses to headwords and to their relevant senses of the paronym by coordinating information in a specific way. As a result, linguistic and extra-linguistic information is more explicit, interlinked, and consistently illustrated, and all entries are harmonised. The four major display elements suitable for contrastive entries are: colour, positioning, sorting principles, and user-generated selection options. They support users in identifying, comparing, and setting new parameters, in changing perspective, and in choosing the relevant parts that are expandable. These functions and modes of presentation and visualisation are not superficial gimmicks, but rather they add valuable information to the descriptions.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

3.2 Language of description and examples

Although the two polysemous terms *legal*/*legitim* both refer to the concept of law (see 2.2), they are not used synonymously, as we can see by analysing actual instances of real language use in corpora. Their individual and distinct contextual uses are therefore placed offset from one another in order to indicate that they are not in a relationship of similarity. Labels that are different enough justify the plausibility of the distinction between senses. Again, the combination of headword, synonym, reference category, and illustrative collocates specifies the contextual environments (Fig. 6).

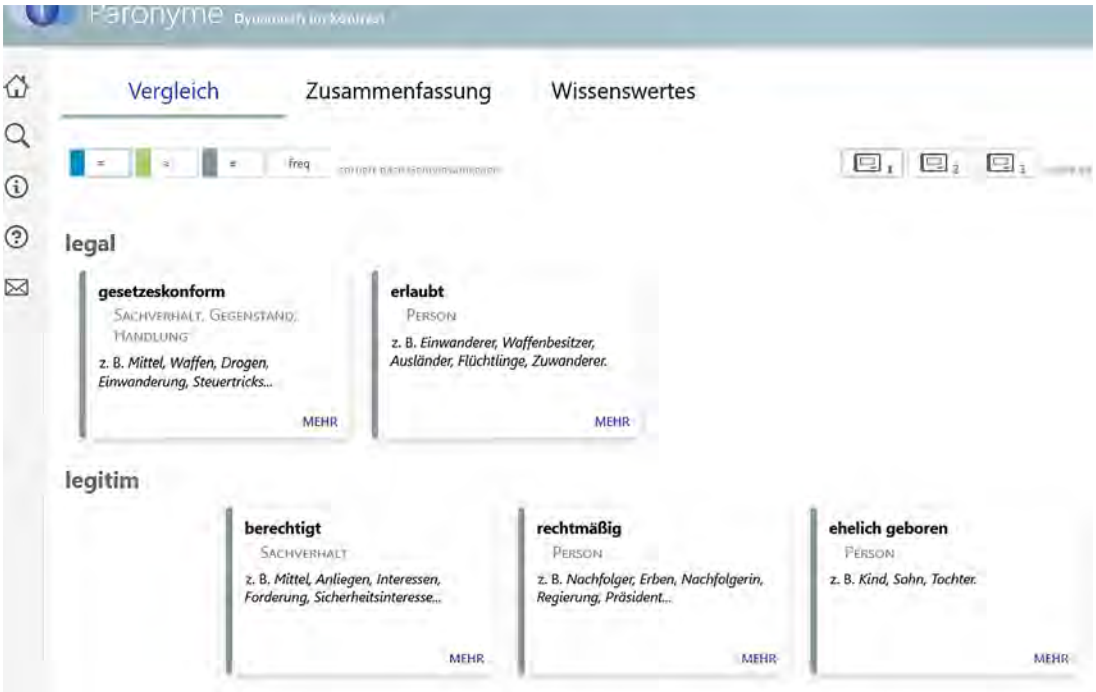


Fig. 6: Entry *legal*/*legitim* in the paronym dictionary

Different lexical information and extra-linguistic details are incorporated into the meaning explanations. These confirm the distinction where two senses of the expressions have been selected (here, two that appear similar at first sight, since they both refer to PEOPLE). The long paraphrase contains a certain amount of entrenched world knowledge. The term *legal* characterises a person in such a way that he/she possesses an official permit or that he/she has proof of a certain official status (e.g. residence permit) or that he/she can prove to be allowed to own specific objects (e.g. weapons). By containing these facts about the real world, the description does not remain abstract but becomes concrete and illustrative. Both adjectives also occur in different thematic domains, which are given beneath the definition (here LAW vs. SOCIETY/CULTURE). As has been pointed out in 3.1, the collocates further illustrate typical syntagmatic constructions as well as realisations of the conceptual reference. People who have been characterised by *legal* can be *Einwanderer* (*migrants*) or *Waffenbesitzer* (*owners of weapons*). Entities which are modified as being *legitim* (*legitimate*) are *Nachfolger* (*successors*), *Erben* (*heirs*), and the *Regierung* (*government*).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Fig. 7: Two distinct senses in the contrastive detail view

Choosing the right corpus examples entails following a number of different criteria. One of them is to have a context where the headword co-occurs together with some of the collocates given above. In the case of equivalent contexts between two usages, the examples must also contain identical patterns (see Fig. 5 *formales/förmliches/formelles Verfahren*). This practice is most effective in providing evidence of collocability, grammatical features, and context-bound near-equivalence with corresponding headwords. By choosing longer definitions with a style closer to everyday language and by avoiding abbreviations altogether, necessary information can be expressed in a more comprehensible manner. The language of the description is more extensive and includes details illustrating and referring to elements of the definition. This approach guarantees a more descriptive and coherent depiction of lexical facts combined with the necessary real-world knowledge.

4. A new dictionary of neologisms (and their synonyms)

The architecture developed for the paronym dictionary is transferable to the description of synonyms in large measure because (near-)synonyms can also cause difficulties as far as their precise differences are concerned. As a next step, we will develop new resources describing German neologisms, including new synonyms such as *Lockdown/Shutdown*, *Corona/Covid/SarsCoV-2* or new loan words with their indigenous counterparts (e.g. *Prank/Streich* (*prank/prank*)). In addition to questions which typically arise for neologistic synonyms, there are similar questions concerning how or whether to distinguish between them. The core feature of neologisms is being new, and therefore they have the potential to be unfamiliar and not yet established in a speaker's mental lexicon. Their assimilation into German might be an ongoing process. Hence, changes as to their adoption of gender, inflectional paradigms, connotations, or even reference are still possible. In Figure 8, a user has a query asking for the difference between the nouns *Covid-19* and *Corona*. S/he provides additional information on an underlying situation that involves seeking details on reference and context: here, the use of both items in terms of a person affected by the disease.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

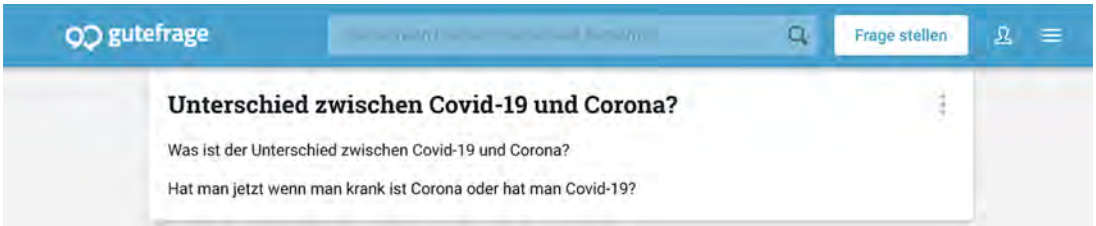


Fig. 8: User asking for the difference between *Covid-19* and *Corona*

In Wiktionary[8] they are both paraphrased as a "disease caused by an infection with Sars-Cov-2", suggesting semantic identity. Fundamental details on the specific reference of the terms are missing, e.g. who is exposed to it, what type of disease it is, and what medical indications or symptoms typically occur. With regard to medical terms, encyclopaedic knowledge is an important part of their semantics and is often sought in queries.

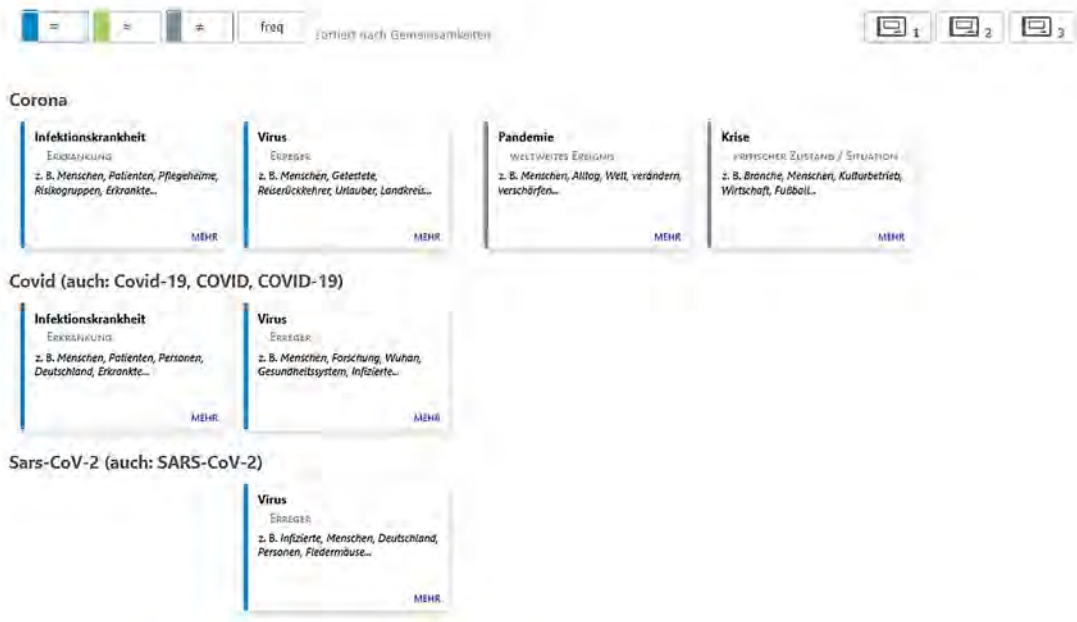


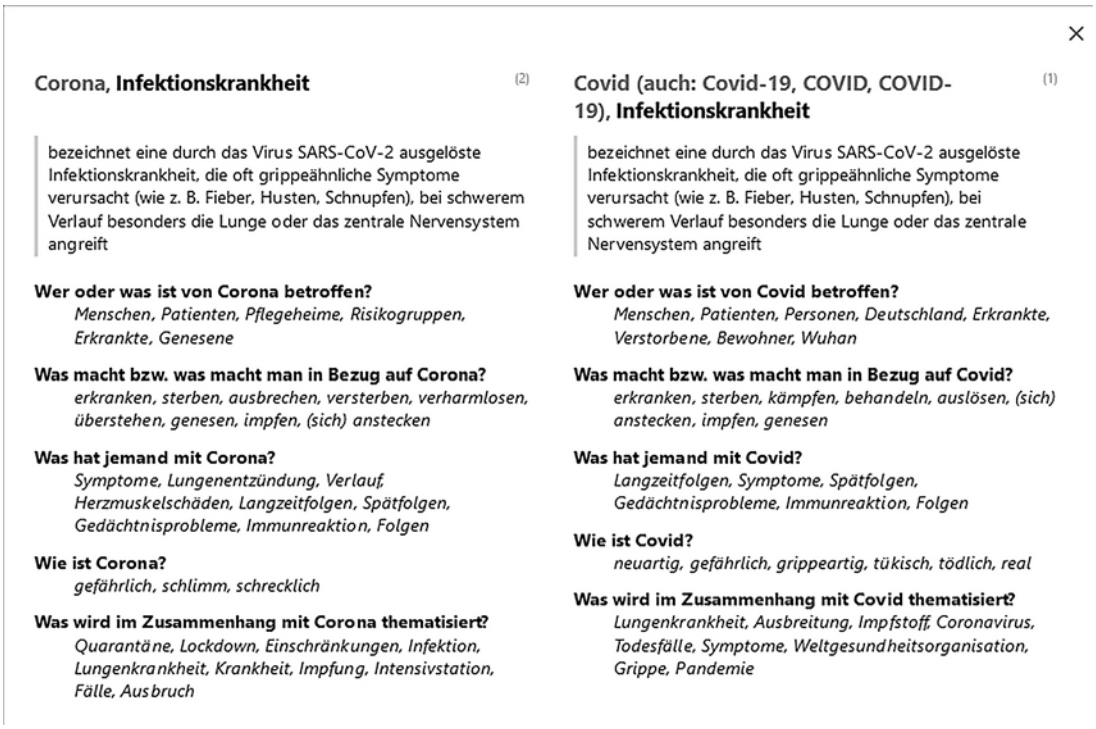
Fig. 9: Neologistic synonyms and collocations

Serving as an example, a fictitious entry including an overview and detailed documentation has been created (Fig. 9) to show the value of a contrastive entry on the basis of the existing paronym dictionary, also illustrating synonymous contexts in everyday language. The information is provided in a similar way, sufficiently disambiguating the senses for each headword. In order to recognise a contextual use, the synonym label ('Infektionskrankheit' or 'Virus') serves as a usage identifier to refer to information relevant for a specific communicative setting. In this case, for *Corona* one context referring to a specific virus and another referring to the infectious disease, a pandemic and a social crisis can instantly be identified. As outlined previously, additional information in both the overview and the detailed view pinpoint particularities concerning who is affected, any activities caused by the virus or

[8] Wiki has a new resource WikiUnterschied.com, published in 2021, which creates comparative entries on the basis of its wiktionary entries (hence similar to table 2). So far, there is no mention of this resource in forums.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

alternatively by the disease), and what is typically associated with it as expressed lexically by collocates.



Corona, **Infektionskrankheit** (2)

bezeichnet eine durch das Virus SARS-CoV-2 ausgelöste Infektionskrankheit, die oft grippeähnliche Symptome verursacht (wie z. B. Fieber, Husten, Schnupfen), bei schwerem Verlauf besonders die Lunge oder das zentrale Nervensystem angreift

Wer oder was ist von Corona betroffen?
Menschen, Patienten, Pflegeheime, Risikogruppen, Erkrankte, Genesene

Was macht bzw. was macht man in Bezug auf Corona?
erkranken, sterben, ausbrechen, versterben, verharmlosen, überstehen, genesen, impfen, (sich) anstecken

Was hat jemand mit Corona?
Symptome, Lungenentzündung, Verlauf, Herzmuskelschäden, Langzeitfolgen, Spätfolgen, Gedächtnisprobleme, Immunreaktion, Folgen

Wie ist Corona?
gefährlich, schlimm, schrecklich

Was wird im Zusammenhang mit Corona thematisiert?
Quarantäne, Lockdown, Einschränkungen, Infektion, Lungenkrankheit, Krankheit, Impfung, Intensivstation, Fälle, Ausbruch

Covid (auch: Covid-19, COVID, COVID-19), **Infektionskrankheit** (1)

bezeichnet eine durch das Virus SARS-CoV-2 ausgelöste Infektionskrankheit, die oft grippeähnliche Symptome verursacht (wie z. B. Fieber, Husten, Schnupfen), bei schwerem Verlauf besonders die Lunge oder das zentrale Nervensystem angreift

Wer oder was ist von Covid betroffen?
Menschen, Patienten, Personen, Deutschland, Erkrankte, Verstorbene, Bewohner, Wuhan

Was macht bzw. was macht man in Bezug auf Covid?
erkranken, sterben, kämpfen, behandeln, auslösen, (sich) anstecken, impfen, genesen

Was hat jemand mit Covid?
Langzeitfolgen, Symptome, Spätfolgen, Gedächtnisprobleme, Immunreaktion, Folgen

Wie ist Covid?
neuartig, gefährlich, grippeartig, tükisch, tödlich, real

Was wird im Zusammenhang mit Covid thematisiert?
Lungenkrankheit, Ausbreitung, Impfstoff, Coronavirus, Todesfälle, Symptome, Weltgesundheitsorganisation, Grippe, Pandemie

Fig. 10: Contrastive sets of collocates between neologistic synonyms

Certainly, the treatment of neologisms needs the kind of detail that typically characterises these terms. Essentially, these relate to the origin, etymology, and morphology of lexemes with foreign elements or the appropriate grammatical use of nouns (e.g. gender, genitive, and plural forms). Inflection paradigms are also important issues for adjectives and verbs. A large number of neologisms also require more discourse-based information (*Shutdown* vs. *Lockdown*) and information about where they first appeared. Nonetheless, the solutions found for the paronym dictionary still seem to serve some needs as far as the comparative aspect of (neologistic) meaning equivalents is concerned (and of synonyms where one term is a loan word). Currently, more studies are being performed looking at different linguistic situations for users for neologisms and loan words and their specific language-related requests, in order to acquire a more complete picture of the new target group.

5. Summary

Language-related deficiencies together with users' dictionary-based problems have not been studied thoroughly in order to improve and design new dictionaries. It is suggested that research on dictionary usage be combined with studies on actual instances of language use. Making adequate distinctions or finding the right word in a specific context when there is more than one option is a frequent subject in chats and blogs. Studying those offers an unprecedented wealth of information about language users, the challenges they face with various dictionaries, and their confusion with paronyms, (near) synonyms, loan words, and neologisms. User studies have assisted dictionary makers in learning about their users, a decisive step forward in building user-friendly resources. However, the insights gained

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

from the investigation of chats had an essential effect on the development of the paronym dictionary. Specifically, they have influenced the contents, presentation, functionality, and style of description. The design and solution-based approach applied in the paronym project centred on gaining a deeper understanding of the target users for whom we actually compile a dictionary (cf. Lew 2015).

Having recognised that users turn to discussion forums, one might wonder whether we still need dictionaries. The answer is "yes" because online forums also tell us about the community's competence, their different intuitions, and their urgent search for reliable reference tools. Personal suggestions vary: often they are limited to prototypical or primary senses, or they are prescriptive, following old educational norms once learned or prevalent in traditional dictionaries. When we scrutinise the target user and his/her linguistic questions before we develop a new product and best combine it with studies of dictionary behaviour and when we redefine lexicographic boundaries and search for new possibilities, we are able to build new dictionaries that are reliable and educational while also being enjoyable to browse through.

References

Antor, H. (1994): Strategien der Benutzerfreundlichkeit im modernen EFL Wörterbuch. In: Henrici, G./Zöfgen, E. (eds.): Fremdsprachen lehren und lernen. Wörterbücher und ihre Benutzer. Tübingen, pp. 65–83.

Chon, Y. V. (2008): The electronic dictionary for writing: a solution or a problem? In: International Journal of Lexicography 22 (1), pp. 23–54.

Dornseiff (2020): Der deutsche Wortschatz nach Sachgruppen. 9th edition. Berlin/Boston.

Duden: www.duden.de (last access: 14-02-2022).

DWDS: Digitales Wörterbuch der deutschen Sprache. https://www.dwds.de/ (last access: 14-02-2022).

elexiko: Online-Wörterbuch zur deutschen Gegenwartssprache. https://www.owid.de/docs/elex/start.jsp (last access: 14-02-2022).

Gutefrage.net. https://www.gutefrage.net/ (last access: 14-02-2022)

Hanks, P. (1987): Definitions and explanations. In: Sinclair, J. (ed.): Looking up – an account of the Cobuild project in lexical computing. London/Glasgow, pp. 116–136.

Leo-Forum (Sprachlabor): https://dict.leo.org/forum/viewGeneraldiscussion.php?idThread=3225&idForum=4&lp=ende&lang=de (last access: 14-02-2022).

Lew. R. (2015): Dictionaries and their users. In: Hanks, P./de Schryver, G.-M. (eds.), International handbook of modern lexis and lexicography. Berlin/Heidelberg, pp. 1–11.

Müller-Spitzer, C. (ed.) (2014): Using online dictionaries. Berlin/Boston.

Murphy, L. (2013): What we talk about when we talk about synonyms (and what it can tell us about thesauruses). In: International Journal of Lexicography 26 (3), pp. 279–304.

Ostermann, C. (2015): Cognitive lexicography. Berlin/Boston.

Paronyme – Dynamisch im Kontrast. https://www.owid.de/parowb/ (last access: 14-02-2022).

Rothe, U. (2014): Das einsprachige Wörterbuch in seinem soziokulturellen Kontext: Gesellschaftliche und sprachwissenschaftliche Aspekte in der Lexikographie des Englischen und des Französischen, Berlin/Boston, pp. 85–113.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Rundell, M. (1999): Dictionary use in production. In: International Journal of Lexicography 12 (1), pp. 35–53.

Storjohann, P. (2018): Commonly confused words in contrastive and dynamic dictionary entries. In: Čibej, J./Gorjanc, V./Kosem, I./Krek, S. (eds.): Proceedings of the XVIII EURALEX International Congress. Lexicography in Global Contexts. Ljubljana, pp. 187–197.

Storjohann, P. (2016): Vom Interesse am Gebrauch von Paronymen zur Notwendigkeit eines dynamischen Wörterbuchs. In: Sprachreport 4/2016, pp. 32–43.

Storjohann, P. (2005): Semantische Paraphrasen und Kurzetikettierungen. In: Haß, U. (ed.): Grundfragen der elektronischen Lexikographie. *elexiko* – das Online-Informationssystem zum deutschen Wortschatz. Berlin/New York, pp. 182–203.

Storrer, A. (2013): Representing dictionaries in hypertextual form. In: Gouws, R./Schweickard, W./Wiegand, H. E. (eds.): Dictionaries. An international encyclopedia of lexicography. Suppl. Vol.: Recent developments with focus on electronic and computational lexicography. Boston/Berlin, pp. 1244–1253.

WikiUnterschied.com. https://wikiunterschied.com/ (last access: 14-02-2022).

WortschatzLeipzig. https://corpora.uni-leipzig.de/de (last access: 14-02-2022).

Contact information

Petra Storjohann
Leibniz-Institut für Deutsche Sprache
storjohann@ids-mannheim.de

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Lexicography: Status, Theory and Methods

XX EURALEX INTERNATIONAL CONGRESS
DICTIONARIES AND SOCIETY
12-16 July 2022, Mannheim, Germany

Konan Kouassi

MENSCH-MASCHINE-INTERAKTION IM LEXIKOGRAPHISCHEN PROZESS ZU LEXIKALISCHEN INFORMATIONSSYSTEMEN

Abstract Dictionaries of today and tomorrow are rather digital products than print dictionaries. From the user's perspective, electronic dictionary applications and in particular „lexical information systems", also referred to as „digital word information systems" are coming to the fore alongside Google searches. Given the rapid developments in the area of the automated provision of lexicographic information, more precisely the automatic creation of online dictionaries, the new role of the lexicographer in the modern lexicographic process is questionable. This article addresses this issue.

Keywords Beispielangaben; computergenerierte Angabe; Funktion des Lexikographen; lexikalisches Informationssystem; redaktionell erstellte Angaben

1. Einleitung

Der vorliegende Beitrag stammt aus dem ProfiLex-Projekt[1] und dient der näheren Beleuchtung der lexikographischen Mensch-Maschine-Interaktion in Hinblick auf die besonderen Einsatzstellen des kognitions- und erfahrungsgeleiteten Lexikographen[2] bei der Erstellung von lexikalischen Informationssystemen. Diese bezeichnen hier aktuelle einsprachige Großwörterbücher des Deutschen, die im Internet frei verfügbar sind, sich an einen breiten Nutzerkreis wenden und deren Bearbeitung zu einem erheblichen Teil automatisch erfolgt. Mittlerweile können, wie Wiegand et al. (2010, S. 15) erwähnen, alle lexikographischen Teilprozesse durch Computertechnik bzw. per Algorithmus unterstützt und betrieben werden. Daher ist vor dem Hintergrund der gegenwärtigen technischen Situation auf Anhieb zu behaupten: Wörterbücher von heute und morgen werden computergeleitet bzw. automatisch erstellt. Es ist aber aufzuklären, was das genau bedeutet. Die einzelnen automatisch erstellten Angaben können korpusgestützt gut oder schlecht gewählt bzw. formuliert sein und stehen deshalb im Mittelpunkt vieler wörterbuchkritischen Abhandlungen (Kirkness 2016; Mollica 2017; Bielińska/Schierholz 2017; Schierholz 2019). Die allgemeine Kritik vertritt die Auffassung, automatisch generierte Angaben seien nicht immer fehlerfrei, und die bemängelten Unzulänglichkeiten sind vielfach. Die wichtigste Frage ist daher mit Recht die konkrete und praktische Rolle des modernen Lexikographen. Der Lexikographische Gesamtprozess besteht projektabhängig aus vielen Arbeitsschritten, so dass eine systematische und exhaustive Antwort nicht gegeben werden kann. Stattdessen wird hier anhand der *Phase der Erstellung und Präsentation von Beispielangaben* (*BeiA*) gezeigt, wie Technik und Mensch systematisch interagieren. Beispielangaben sind u. a. Bereiche, in denen diese Inter-

[1] ‚Der Professionelle Lexikograph', kurz ‚ProfiLex-Projekt' ist ein metalexikographisches und wissenschaftliches Projekt, das am Interdisziplinären Zentrum für Lexikografie, Valenz- und Kollokationsforschung der FAU Erlangen-Nürnberg (URL: http://www.lexi.uni-erlangen.de/de/) geführt wird. Kernpunkt des Projekts besteht in der Erforschung der Funktionen und Aufgaben des professionellen Lexikographen in Abgrenzung zu den automatischen Verfahren im modernen lexikographischen Prozess für lexikalische Informationssysteme.

[2] Im Text wird das generische Maskulinum verwendet und bezeichnet so Personen aller Geschlechter.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

aktion sehr zielgenau beschrieben und aufgezeigt werden kann. Es werden drei Fragestellungen verfolgt: (i) Wo treffen Technik und Lexikograph bei der Erstellung von Beispielangaben aufeinander? (ii) Welche Merkmale kennzeichnen dabei sowohl den technischen als auch den intellektuellen Prozess? (iii) Ist ein vollautomatisches Verfahren qualitativ gesehen sinnvoll? Erhoben wird die Problematik des automatischen Datenanbietens für Benutzer, die auf eine Sach- oder Sprachfrage eine korrekte und vollständige Antwort erwarten.

Der Beitrag ist inhaltlich in zwei Hauptteile gegliedert: Im ersten Schritt wird der Erstellungsprozess von Beispielangaben zu lexikalischen Informationssystemen dargelegt. Der darauffolgende Teil befasst sich mit dem Kontrast zwischen redaktionell und automatisch erstellten Beispielangaben.

2. Der Erstellungsprozess von Beispielangaben zu lexikalischen Informationssystemen

In vielen Wörterbüchern schließen sich die Beispielangaben (BeiA) den Bedeutungsparaphrasenangaben an und dienen u. a. zur Erläuterung und Veranschaulichung der Bedeutungsangaben, zur Illustration des behandelten Lemmazeichens bzw. zur Präsentation von syntaktischen Eigenschaften und Verwendungsweisen (vgl. Lettner 2020, S. 32). Wie Haß-Zumkehr (2001, S. 35) schreibt, seien in der einsprachigen Lexikographie die Beispielangaben „neben der Bedeutungsangabe in irgendeiner ihrer Formen [ob Kompetenz-, Beleg- oder Korpusbeispielangaben] wohl die wichtigste und eigenständig wahrgenommene Angabeart“. Atkins/Rundel (2008, S. 453) präzisieren schon: „Attaching examples to definitions is a separate process“. Doch ein Beispiel kann gut oder schlecht gewählt bzw. formuliert sein. Darum sind bei dieser Aufgabe unabhängig vom Beispieltyp, vom Erstellungs- und Präsentationsprozess der intellektuelle Einsatz und eine besondere Geschicklichkeit seitens des Lexikographen gefordert. Im Weiteren wird davon ausgegangen und gezeigt, dass für digitale lexikalische Informationssysteme zutreffende Beispielangaben vor dem Hintergrund der angelegten Wörterbuchfunktionen und -adressaten erst durch die Komplementarität bzw. Interdependenz von Lexikographenexpertise und Technik zuverlässig entstehen können.

2.1 Die Automatik bei der Beispielerstellung für Wörterbuchartikel

Die spezifischen lexikographischen Tätigkeiten sind heute auf technische und technologische Mechanismen angewiesen, welche immer weiter verbessert werden. Zur Erstellung von Beispielangaben beginnt der Prozess bereits bei der linguistischen Datenaufbereitung zur Wörterbuchbasis. Wie die Daten aufbereitet werden, bestimmt, wie sie in den weiteren lexikographischen Teilprozessen zu verwenden und auszuwerten sind. Dabei ist zunächst einmal aus funktionalen Gesichtspunkten zu unterscheiden zwischen Beispielformaten[3] zum Zwecke des Korpusaufbaus und solchen zum Zwecke der eigentlichen Erstellung und Präsentation von Beispielangaben (vgl. Lenz 1998). Beispiel- bzw. Belegangaben hätten zum Korpusaufbau für den Computer eine wörterbuchbasisbezogene Angabefunktion, aber auch

[3] Die aufbereiteten Primärquellendaten zur Wörterbuchbasis enthalten Belege bzw. Stichwortokkurrenzen und Stichwortprofile und dienen zur Attestation oder Dokumentation des potenziellen Stichwortes im Kontext.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

zur konkreten Beispielerstellung für die Benutzer eine wörterbuchgegenstandsbezogene Angabefunktion (vgl. Wiegand 2006, S. 269). Die Automatik bei der Beispielerstellung setzt voraus, dass der Computer als Arbeitsmittel mit bestimmten technischen Infrastrukturen ausgestattet ist (vgl. Abel/Klosa 2012, S. 414 f.). Dazu gehören u. a. Artikelredaktionssystem, Korpusabfragesystem, Datenbank, Belegextraktionsprogramme, die zur Unterstützung und Optimierung der lexikographischen Arbeitsprozesse eingesetzt werden (vgl. Klosa/Tiberius 2018, S. 96). Etliche Arbeitsschritte wie Sammeln, Ordnen und Sortieren laufen automatisch schnell und präzise durch die Computerprogramme. Automatische Verfahren sind auch sinnvoll, wenn umfangreiches sprachliches Material erhoben, analysiert und modifiziert wird. Dies betrifft Korpusanalysen zur Frequenz, Kookkurrenz und N-Gramm-Verfahren (vgl. u. a. Klosa 2007; Prinsloo 2009). Dazu gehören auch automatische sprachliche Korrekturen (Rechtschreibung, Grammatik), Datenpräsentation u. v. m. Automatische Verfahren basieren meistens auf der formalen Seite sprachlicher Zeichen. Und gerade vor dem Hintergrund einer instabilen und dynamischen Wörterbuchbasis, die im Rahmen von lexikalischen Informationssystemen ständig einer Erweiterung und Aktualisierung unterzogen wird, vor allem, wenn die Beispielangaben die Wörterbuchbasis widerspiegeln sollen, ist eine automatische Abstraktion von Beispielangaben zwar möglich, aber unkontrollierbar. Denn es entstehen vor allem semantische Ambiguitäten, die den speziellen Einsatz des Lexikographen erfordern.

2.2 Der intellektuelle Einsatz des Lexikographen

Der intellektuelle Einsatz begleitet den gesamten Beispielerstellungsprozess und geht eigentlich durch Konzeptionen, Implementierungen und kritische Beobachtungen des automatischen Beispielerstellungsverfahrens oder schließt sich diesem an. Lexikographiegeschichtlich werden in den meisten Wörterbuchprojekten die Beispielangaben nicht automatisch erstellt. Und für Bedeutungswörterbücher sind Beispielangaben obligatorische Angaben. Insbesondere die Kompetenzbeispielangaben (KBeiA) werden anhand eines (möglicherweise) vorliegenden Instruktionsbuchs und auf der Basis der technischen Infrastruktur mit der Idiokompetenz des Lexikographen gebildet (vgl. Kunze/Lemnitzer 2007, S. 82). Der Einsatz des Lexikographen besteht darin, dass er mit seinen eigenen intellektuellen Kompetenzen und Erkenntnissen Wege sucht, wie er angemessene, verständliche, hilfreiche KBeiA präsentiert; er trägt somit die volle Verantwortung für die Qualität der Angaben. Erforderlich ist dabei die persönliche Einsicht in die semantischen Eigenschaften des zu behandelnden Lemmas. Das Bedeutungsparadigma und die Bedeutungsbeschreibung sind somit die unmittelbare Voraussetzung für die Erstellung und Zuordnung von KBeiA. Dafür ist zunächst die Sichtung der automatisch generierten Belege durch Computeralgorithmen (Konkordanz-, Belegextraktionsprogramme) erforderlich, indem basierend auf der sprachlichen Form Korpusbelege aufgerufen werden. Bei der Ermittlung der semantischen Eigenschaften, der Identifizierung von relevanten Belegen und deren Verbeispielung bzw. Transformation zu Beispielangaben ist die Expertise des professionellen Lexikographen gefordert. Für die Bildung von Kompetenzbeispielen kann er aus den Belegen ein Textsegment, einen Satz oder eine Wortgruppe als angemessen für eine Beispielangabe übernehmen sowie sich auf Belege stützend und entsprechend den angelegten Wörterbuchfunktionen und potenziellen Adressaten eigene und pragmatische Beispiele erstellen. Diese können vollständige Einzelsätze oder Satzgruppen, typische Wortverbindungen, Kollokationsbildungen, Infinitivkonstruktionen, paradigmatische Ausdrücke, kurz oder lang sein (vgl. Lettner 2020, S. 161), so dass der Lexikograph vor dem Hintergrund der anvisierten Benutzer mehr Frei-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

heiten und Chancen hat, die tatsächlichen Benutzerbedürfnisse zu treffen. Soweit möglich, sollten die erstellten KBeiA an den gefundenen Korpusbelegen orientiert sein.

Ein weiterer Schritt ist die Benutzerbezogenheit, was voraussetzt, dass durch Kenntnis der Adressaten und deren Voraussetzungen zur Wörterbuchkonsultation passende Beispiele erstellt werden können. Somit muss der Lexikograph relevante Belege erkennen, Wörter und Ausdrücke für die Beispielangaben wählen, deren sprachliche Form sowie die Anzahl und Reihenfolge der Beispielangaben bestimmen, die Korrelation zwischen Lemmazeichen, Bedeutungsbeschreibung und Beispielangaben ermessen. Diese Entscheidungsschritte sind unumgänglich und können nicht zuverlässig durch Computerprogramme erreicht werden. Letztere können lediglich wortformbasierte Analysen ausführen und Daten aus anderen Quellen übernehmen.

Im Umgang mit authentischen Sprachdaten als Wörterbuchbasis, die die gesellschaftliche Denkweise (der Wörterbuchsprache) widerspiegeln, ist es notwendig, dass der Lexikograph nicht nur diese gesellschaftliche Denkweise erkennt und beherrscht, sondern auch die Sprache, die damit in Verbindung steht. Es hat damit zu tun, dass der Lexikograph diese gesellschaftliche Denkweise auch in den Beispielangaben widerspiegeln lassen kann. Er erwirbt dies als Muttersprachler oder durch Erlernen der behandelten Sprache und Kultur.

Festzuhalten hat also der Einsatz des Lexikographen bei der Erstellung von KBeiA folgende Ausprägungen:

a) Festlegung und Einhaltung des Beispielkonzepts.

b) Adäquater Umgang mit der angelegten technischen Infrastruktur.

c) Kognitionsgeleitete Erkennung von wörterbuchgegenstangsbezogenen relevanten Belegen.

d) Intellektuelle Berücksichtigung von Wörterbuchfunktion und potenziellen Adressaten.

e) Intellektbasierte Erschließung des semantischen Spektrums zum Lemma.
(Punkt *a* bis *e* gelten als Vorfeldschritte zur eigentlichen Erstellung von KBeiA)

f) Wahl der lexikalischen Einheiten bzw. Konversion/Transformation der Vorfeldschritte zu konkreten KBeiA.

g) Intellektbasiertes Ermessen von Beispielqualität.

Diese Arbeitsschritte und -prozesse können zum heutigen Stand der Technik nicht durch Computer erreicht werden.

3. Beispielangaben in digitalen Wortauskunftssystemen

Es werden hier Beispielangaben im Wörterbuch *Duden online* und *DWDS* als lexikalische Informationssysteme näher betrachtet. Es geschieht dabei eine Art Zurückverfolgung der Entstehung der Beispielangaben, um die jeweilige Beteiligung von Lexikograph und Computertechnik aus einer anderen und metalexikographischen Perspektive zu beobachten. Der erste Fall handelt von redaktionell erstellten BeiA; der zweite von automatisch erstellten BeiA.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

3.1 Redaktionell erstellte BeiA am Beispiel von *Duden online*

Gegeben sei der Beispielkomplex zur ersten Unterbedeutung der ersten Bedeutung zum Lemma **Gedanke,** [in Abb. 1 realisiert durch „**1.a)**“] (vgl. Abb. 1).[4]

In welcher Abhängigkeitsrelation Bedeutungsangabe und Beispielkomplex stehen, bzw. was erst erstellt wird, ist zwar über die Umtexte nicht erschließbar. Landau (2004, S. 210) schlägt im Zusammenhang mit redaktionell ertellten BeiA vor: „Using invented examples is like fixing a horse race: the lexicographer invents an example to justify his definition instead of devising a definition to fit the examples”. Dabei hebt Landau die Authentizität und besondere Bedeutung von Korpusbeispielen hervor. Lexikographiegeschichtlich geschieht Landaus Ansatz aber nur zum Zwecke der Erschließung des Bedeutungsparadigmas des zu bearbeitenden Lemmas durch intellektuelle Korpusbeleganalysen, sodass die angesetzten Beispielangaben zur Begründung, Illustration und Weitererklärung der Bedeutungsangabe dienen. Die Beispielangaben sind deshalb auch von den spezifischen Lesarten abhängig. Abbildung 1 weist einen quantitativ hybriden Beispielkomplex auf, was auf die freie Entscheidung des Lexikographen zurückgeführt werden kann. Die Analyse der weiteren Beispielkomplexe (zu den insgesamt fünf Hauptbedeutungen) desselben Lemmas sowie anderer Wörterbuchartikel gibt dazu eine Bestätigung (Menge und Struktur sind pauschal bestimmt). Die Beispielangaben in Abbildung 1 sind teilweise Kollokationen (z. B.: *gute, vernünftige Gedanken*), Ganzsatzbeispiele (z. B.: *ein Gedanke ging mir durch den Kopf*), einzelne Beispielgruppen (z B.: *dieser Gedanke liegt mir fern, verfolgt mich, tröstet mich*), Infinitivkonstruktionen (z. B.: *auf einen Gedanken kommen*), oft mit kursiv gesetzten pragmatischen Markierungen (z. B.: „seine Gedanken sammeln (*sich konzentrieren*)“) etc.

Gedanke 1. a) etwas, was gedacht wird, gedacht worden ist; Überlegung

BEISPIELE

- gute, vernünftige Gedanken
- dieser Gedanke liegt mir fern, verfolgt mich, tröstet mich
- ein Gedanke ging mir durch den Kopf
- mir drängt sich der Gedanke auf, dass das nicht stimmt
- einen Gedanken fassen, aufgreifen, fallen lassen, in Worte kleiden, zu Ende denken, nicht mehr loswerden
- Gedanken an jemanden, etwas verschwenden
- auf einen Gedanken kommen, verfallen
- es ist mir ein schrecklicher Gedanke (*eine schreckliche Vorstellung*), dass du verärgert bist
- seine Gedanken sammeln (*sich konzentrieren*)
- seinen Gedanken nachhängen, sich seinen Gedanken überlassen ([*nach*]*sinnen*)
- (...)

Abb. 1: (Verkürzter) Beispielkomplex zum Lemma **Gedanke** im *Duden online*

[4] Lemma **Gedanke** im *Duden online*: https://www.duden.de/rechtschreibung/Gedanke (Stand: 10.5.2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Aus der Fülle der Korpusbelege und vor dem Hintergrund der erschlossenen Lesarten muss der Lexikograph zum Beispiel die *Entscheidungsfunktion des Lexikographen* anwenden, um BeiA im Zusammenhang mit Lesart und Benutzerbedürfnis für den spezifischen Wörterbuchartikel zu erstellen. Funktion heißt hier, wozu der Lexikograph da ist oder sein soll. Außerdem wendet er die *Interpretationsfunktion* bei der semantischen Analyse der Korpusbelege an und muss die *Redaktionsfunktion* (eigene Sprachfähigkeit in Zusammensetzung mit Benuterbezogenheit) ebenso heranziehen. Am Ende tätigt er die *Validationsfunktion.* Diese ist der intellektuelle Prozess, durch den Beispielangaben durch den Lexikographen ermessen und als korrekt erkannt und angesetzt werden. Es geschieht also allein bei der Erstellung von KBeiA der Einsatz von mehreren Funktionen des Lexikographen. Das Fehlen solcher Leistungen durch den Lexikographen erzeugt Ergebnisse wie im folgenden Beispiel.

3.2 Automatisch erstellte BeiA am Beispiel des DWDS-Modells

Die automatisch generierten Korpusbeispiele unter der Rubrik „Verwendungsbeispiele“ im DWDS[5] stehen auch als Beispielangaben und gelten als Blickwinkel für die Interaktion oder Abgrenzung der Computerleistungen. Die Annahme, dass der Lexikograph bei semantischen Angelegenheiten dem Computer besonders überlegen ist, lässt sich vor allem an der automatischen Bearbeitung von homographen Lemmata beobachten. Angenommen seien zum Beispiel die automatisch generierten Verwendungsbeispiele zum homographen Lemma **Verband**.[6] Dort werden unter „Verwendungsbeispiele“ die folgenden fünf Belege automatisch angezeigt:

- Im vergangenen Jahr schätzte der **Verband** ihre Zahl auf 284.000.
- Unterstützt werden die **Verbände** von etwa 2 Millionen ehrenamtlichen Helfern.
- Die verschiedenen **Verbände** gehen sich aber meistens aus dem Wege.
- Mittlerweile hatte der Allgemeine **Verband** eine andere Taktik für zweckmäßig erachtet.
- Als sie den **Verband** angelegt hatte, griff er nach ihrer Hand und küsste sie dankbar.

Das homographe Lemma **Verband** findet man in der Bedeutung-1 ‘Wundverband’ (Abb. 2) und in der dreifach-polysemen Bedeutung-2 (Abb. 3).

Verband1

Bedeckung einer Wunde, kranken Stelle am Körper durch Verbandsmaterial zum Schutz gegen Infektion und zur Förderung der Heilung

Abb. 2: Ausschnitt des semantischen Kommentars zu Verband1 aus dem DWDS

[5] *Digitales Wörterbuch der Deutschen Sprache* (DWDS): https://www.dwds.de/.

[6] DWDS: https://www.dwds.de/wb/Verband#1 (Stand: 10.3.2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Verband2

1. Verbindung
   a) [Geologie] von Erzen oder Kohle mit dem umgebenden Gestein, in das sie eingebettet sind
   b) [Bauwesen] von Baumaterial, besonders Ziegelsteinen oder Balken, die beim Bauen über Fugen verlegt werden, sodass sie sich gegenseitig stützen und dem Gebäude größere Stabilität geben
   c) von den einzelnen Fasern eines Gewebes
2. Zusammenhang, Gruppierung
   a) von Menschen oder Tieren
   b) [Militär] Vereinigung mehrerer Truppenteile einer oder verschiedener Waffengattungen
3. Organisation, Vereinigung von Menschen, Menschengruppen zur Wahrung und Durchsetzung gemeinsamer Ziele und Zwecke, Bund, Gesellschaft

Abb. 3: Ausschnitt des semantischen Kommentars zu Verband2 aus dem DWDS

Dieselben Belege werden sowohl für Verband1 als auch für Verband2 generiert und sind dabei nicht den Homographen zugeordnet. Das letzte Beispiel gehört zu Verband1, die übrigen zu Verband2. Hier wird die Polysemie nicht beachtet. Die Zuordnung müssen die Benutzer selbstständig leisten. Es fehlt der Einsatz des professionellen Lexikographen; denn im DWDS gibt es Belege, die man den Homographen bzw. der Polysemie zuordnen kann. Die Analysen weiterer Wörterbuchartikel zu Homographen ergeben vergleichbare Resultate.

Die automatische Generierung von Beispielangaben ergibt also ein quantitativ reiches Angebot, wie es zu lexikalischen Informationssystemen gehört, das aber qualitativ arm ist. Die Daten suggerieren eine Vielfalt des lexikalischen Informationssystems, die für den Benutzer keinen Gewinn erbringt. Für ein hochqualifiziertes Sprachdatenangebot ist eine wesentlich sorgfältigere Analyse und Interpretation erforderlich. Die Analyse zur lexikographischen Bearbeitung homographer Lemmata zeigt, wie schnell die Extraktion-Software an Grenzen stößt. Die Benutzer können durch die automatisch generierten lexikographischen Daten irritiert werden oder bemerken die Fehler gar nicht und entnehmen den Daten falsche Informationen. Vom Benutzer werden zudem stillschweigend hohe Interpretationsleistungen zum Datenangebot eingefordert. Da die automatische Sprachdatenauswertung und maschinelle Produktion lexikographischer Daten auf die Wortform begrenzt sind, bleiben semantische Eigenschaften unberücksichtigt. Im Beispielfall werden die Belege methodisch vom DWDS-Beispielextraktor nach sogenannten globalen und lokalen Kriterien[7] automatisch ausgesucht. Bei näherem Hinsehen weisen die Kriterien lediglich formale Motivationen auf; semantische Motivationen können nicht eingeschlossen werden, sodass der Lexikograph mit intellektueller Kompetenz Einsatz machen muss (vgl. 2.2). Es ist deshalb recht, dass zwar formale Aspekte automatisch erfolgen, die Inhaltsseite sprachlicher Ausdrücke aber durch den professionellen Lexikographen mit seinen Interpretations-, Beurteilungs- und Entscheidungskompetenzen sowie Funktionen zur Erreichung hochqualifizierter und zuverlässiger Datenangebote bearbeitet wird.

[7] DWDS: https://www.dwds.de/d/beispielextraktor: dort unter **‚Methode'** (Stand: 10.5.2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

4. Fazit

BeiA können zur Beleuchtung der Interaktion zwischen Mensch und Computertechnik im modernen lexikographischen Prozess für einsprachige Online-Wörterbücher des Deutschen herangezogen werden. Die Ausführungen zum Beitrag der Computertechnik und des Lexikographen zeigen, dass beide Komponenten bei der Erstellung von BeiA bzw. KBeiA interagieren. Der Computer mag zwar automatische Analysen sowie Extraktionsprozesse so durchführen, wie kein Mensch es leisten kann. Mensch und Maschine treffen sich also dort, wo die Erledigung einer bestimmten lexikographischen Aufgabe dem Menschen unmöglich oder zu schwer und langwierig realisierbar ist, oder die Aufgabe sich vom Computer allein nicht zuverlässig bearbeiten lässt. Automatische Verfahren sind sinnvoll und sicher vor allem bei den lexikographischen Angaben des Formkommentars. Ein sicheres automatisches Datenangebot bei semantischen Aspekten erfordert daher beide Komponenten als obligatorisch und in einem Komplementaritätsverhältnis. Es bleibt noch offen, bis zu welchem Umfang der jeweilige Anteil von Mensch und Maschine zu überschlagen ist.

Literatur

Abel, A./Klosa, A.e (2012): Der lexikographische Arbeitsplatz – Theorie und Praxis. In: Fjeld, R./Torjusen, J. M. (Hg.): Proceedings of the 15th EURALEX International Congress. Oslo, S. 413–421.

Atkins, B. T. S./Rundel, M. (2008): The Oxford guide to practical lexicography. Oxford/New York.

Bielińska, M./Schierholz, S. (Hg.) (2017): Wörterbuchkritik – Dictionary criticism. Berlin/Boston.

Duden-Redaktion: Duden online: https://www.duden.de/woerterbuch (Stand: 10.5.2022).

Haß-Zumkehr, U. (2001): Deutsche Wörterbücher – Brennpunkt von Sprach- und Kulturgeschichte. Berlin.

Kirkness, A. (2016): Es leben die Riesenschildkröten! Plädoyer für die wissenschaftlich-historische Lexikographie des Deutschen. In: Lexicographica 32, S. 17–137.

Klosa, A. (2007): Korpusgestützte Lexikographie: besser, schneller, umfangreicher? In: Kallmeyer, W./Zifonun, G. (Hg.): Sprachkorpora – Datenmengen und Erkenntnisfortschritt. Berlin/New York, S. 105–122.

Klosa, A./Tiberius, C. (2018): Der lexikographische Prozess. In: Klosa, A./Müller-Spitzer, C. (Hg.): Internetlexikografie. Ein Kompendium. Berlin, S. 65–110.

Kunze, C./Lemnitzer, L. (2007): Computerlexikographie. Eine Einführung. Tübingen.

Landau, S. I. (2004): Dictionaries. The art and craft of lexicography. 2. Auflage. Cambridge.

Lenz, A. (1998): Untersuchungen zur Beispiel- und Beleglexikographie historischer Bedeutungswörterbücher unter besonderer Berücksichtigung der Neubearbeitung des Deutschen Wörterbuchs gegründet von Jacob und Wilhelm Grimm.
[Online unter: https://ediss.uni-goettingen.de/handle/11858/00-1735-0000-000D-F20D-0 (Stand: März 2022).

Lettner, K. (2020): Zur Theorie des lexikographischen Beispiels. Berlin/Boston.

Mollica, F. (2017): Wörterbuchkritik und Wörterbuchbenutzungsforschung. In: Bielińska, M./Schierholz, S. (Hg.): Wörterbuchkritik – Dictionary criticism. Berlin/Boston, S. 133–171.

Prinsloo, D. (2009): The role of corpora in future dictionaries. In: Nielsen, S./Tarp, S. (Hg.): Lexicography in the 21st century. In honour of Henning Bergenholtz. Amsterdam/Philadelphia, S. 181–206.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Schierholz, S. J. (2019): Brauchen wir noch Wörterbücher? – Ja! – Aber welche? In: Eichinger, L./Plewnia, A. (Hg.): Neues vom heutigen Deutsch. Empirisch – Methodisch –. Berlin/Boston, S. 163–198.

Wiegand, H. E. (2006): Angaben, funktionale Angabezusätze, Angabetexte, Angabestrukturen, Strukturanzeiger, Kommentare und mehr. Ein Beitrag zur Theorie der Wörterbuchform. In: Lexicographica 21, S. 202–379.

Wiegand, H. E./Beißwenger, M./Gouws, R./Kammerer, M./Storrer, A./Wolski, W. (2010): Systematische Einführung. In: Wiegand, H. E. et al. (Hg.): Wörterbuch zur Lexikographie und Wörterbuchforschung. Mit englischen Übersetzungen der Umtexte und Definitionen sowie Äquivalenten in neun Sprachen. Berlin/New York, S. 1–105.

Contact information

Konan Jean Mermoz Kouassi
Friedrich-Alexander-Universität Erlangen-Nürnberg
konan.kouassi@fau.de

Acknowledgements

ProfiLex-Projekt „Gefördert durch die Deutsche Forschungsgemeinschaft (DFG) - 494892119“.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Ana Salgado/Rute Costa/Toma Tasovac

APPLYING TERMINOLOGICAL METHODS TO LEXICOGRAPHIC WORK: TERMS AND THEIR DOMAINS

Abstract Applying terminological methods to lexicography helps lexicographers deal with the terms occurring in general language dictionaries, especially when it comes to writing the definitions of concepts belonging to special fields. In the context of the lexicographic work of the *Dicionário da Língua Portuguesa*, an updated digital version of the last Academia das Ciências de Lisboa' dictionary published in 2001, we have assumed that terminology – in its dual dimension, both linguistic and conceptual – and lexicography are complementary in their methodological approaches. Both disciplines deal with lexical items, which can be lexical units or terms. In this paper, we apply terminological methods to improve the treatment of terms in general language dictionaries and to write definitions as a form of achieving more precision and accuracy, and also to specify the domains to which they belong. Additionally, we highlight the consistent modelling of lexicographic components, namely the hierarchy of domain labels, as they are term identification markers instead of a flat list of domains. The need to create and make available structured, organised and interoperable lexicographic resources has led us to follow a path in which the application of standards and best practices of treating and representing specialised lexicographic content are fundamental requirements.

Keywords Definition; domain label; general language dictionary; lexicography; term; terminology

1. Introduction

The title of this paper highlights our belief that terminology as a science with its own methodology and interdisciplinary and transdisciplinary nature (Felber 1987, p. 1) can contribute to a practice-based rethinking of lexicographic work when terms are at the core of the analysis. We will demonstrate on these pages that terminological methods can help lexicographers and are advantageous for the process of lexicographic knowledge-building.

Due to the democratisation of knowledge, the growth of communication media and the technological and scientific boom, terms are exceptional sources of lexical renewal and enrichment of the language systems. Thus, their registration in general language dictionaries has increased over the years (Rondeau 1984, pp. 1–4).

Many researchers have conducted studies on the presence of terms in general language dictionaries based on monolingual dictionaries (Rey 1985; Béjoint 1988; Tournier 1992; Cabré 1994; Paz Battaner 1996; Estopà 1998; Boulanger 2001; Roberts 2004; Guerra Salas/Gómez Sánchez 2005; Nomdedeu Rull 2008), reviewing different topics (e. g., coverage and percentage of terms, domain labelling, terms related to a specific domain, etc.). We distance ourselves from these authors whenever we apply terminological methods to lexicographic work since we believe that lexical units (words in general) and terminological units (terms) must be differentiated. Lexicography and terminology are two disciplines with different theoretical and methodological assumptions and whose final products aim to respond to different social needs. In this context, we will describe the method we apply to treat terms in general language dictionaries, mainly based on International Organisation for Standardisation (ISO) standards, namely ISO 704 (2009) and ISO 1087 (2019).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

In the universe of the labelling system commonly used in lexicography, labels assigned to specialised senses are called domain labels, a 'marker which identifies the specialised field of knowledge in which a lexical unit is mainly used' (Salgado/Costa/Tasovac 2019). These markers represent the most efficient method to detect terms in general language dictionaries, which justifies our interest in this type of label.

For the sake of consistency, throughout this paper, we have adopted some typographic conventions as exemplified below:

- Domain labels are written in small caps, e. g., GEOLOGY.
- Terms are written in quotation marks, e. g., "term".
- Concepts are written in angled brackets and with the first letter capitalised in a fixed-width (monospace) font, e. g., `<Concept>`.
- Concept relation identifiers are written with an underscore between the forms in a fixed-width (monospace) font, e. g., `is_a`.
- TEI P5[1] terms are written in a fixed-width (monospace) font.

The rest of this paper is organised as follows. Section 2 aims to start by clarifying some of the key concepts, namely term, which necessarily brings concept along, and, subsequently, we conduct our research in light of the double dimension of terminology. Section 3 presents our dictionary case study and the domain selected for the study (GEOLOGY). Section 4 is dedicated to the applied terminological working methods used in a Portuguese language dictionary. Finally, we present concluding remarks and highlight our future work.

2. Framework issues

Term and concept are two core keywords that have been defined quite differently by the various theoretical approaches in terminology (e. g., Wüster 1979/1998; Felber 1987; Cabré 1999; Temmerman 2000; Gaudin 2007; Faber 2009). Despite that, we adopted ISO definitions, i. e., a *term* is understood as a '*designation* that represents a *general concept* by linguistic means' (ISO 1087 2019, p. 7), and a *concept* 'should be viewed not only as a unit of thought but also as a unit of knowledge' (ISO 704 2009, p. 3) 'created by a unique combination of *characteristics*' (ISO 1087 2019, p. 3). In other words, the *concept* – a non-linguistic element – is designated by the term, and the *term* – a linguistic element – in turn lexically designates the concept (Fig. 1).

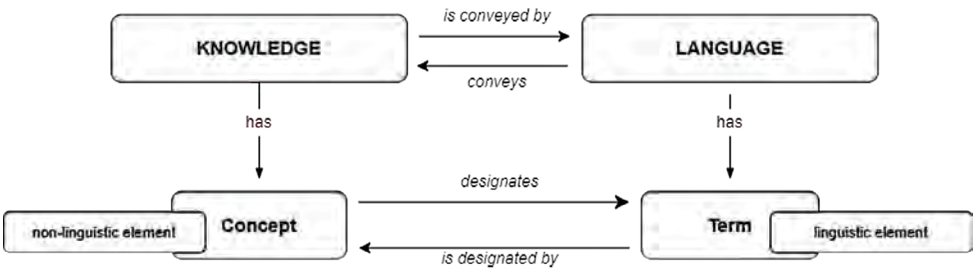


Fig. 1: The Relationship of Concept and Term mirroring the double dimension of terminology (adapted from Costa 2021)

[1] https://tei-c.org/guidelines/p5/.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

We always bear in mind that terms lexically designate concepts, which are often not of primary concern to lexicographers, who usually start from the word form to identify senses, pushing the concept to a secondary level, or ultimately disregarding it. Instead of following this semasiological approach, we propose a different and combined perspective prioritising the concept.

Since the concept 'is created by a unique combination of *characteristics*' (ISO 1087 2019, p. 3) we need to know that a *characteristic* is an 'abstraction of a *property*' (ISO 1087 2019, p. 3). We have paid attention only to the so-called *essential characteristics* – '*characteristic* of a *concept* that is indispensable to understand that concept' (ibidem). As we will see, the distinctive characteristics of a concept are fundamental for creating concept systems and drafting definitions.

Throughout this work, we analyse the terms anchored in the double dimension of terminology (Costa 2013; Roche 2015), where we will reconcile iteratively, step by step, both the onomasiological and semasiological approaches. The onomasiological perspective makes us look at the concept designated by the term, identify it, isolate it, specify its characteristics to differentiate it from other concepts that belong to the same concept system. Finally, the concept is embedded in the concept system where it belongs. This approach is complemented by the traditional lexicographic methodology, which follows a semasiological path, in the sense that it begins from an existing corpus of specialised lexical units (the terms collected from the dictionary that will be referred to in the next section) to explore their semantic values. Following this mixed approach, only after the relations are well-established, will the lexicographer be able to propose a definition that must be validated by the domain expert. We have aimed to combine the conceptual perspective – i. e., knowledge organisation – with the linguistic perspective – focusing on the terms themselves by analysing the data extracted from the dictionary under study.

3. *Dicionário da Língua Portuguesa* as a case study

Our methodology has been applied to a scholarly dictionary of the Portuguese language now being developed by the Academia das Ciências de Lisboa – the *Dicionário da Língua Portuguesa* (DLP) (ACL 2021). This lexicographic work is a retro-digitised dictionary (Simões/Almeida/Salgado 2016) whose starting point was the *Dicionário da Língua Portuguesa Contemporânea* (DLPC) (ACL 2001), last published in 2001. Currently, it is being prepared under the Instituto de Lexicologia e Lexicografia da Língua Portuguesa's supervision in collaboration with researchers and invited collaborators. This project is supported by a small annual Community Support Fund Portuguese National Fund (Fundo de Apoio à Comunidade – FAC) through the Fundação para a Ciência e a Tecnologia (FCT). It will be the first academy Portuguese digital dictionary and it will soon be available online.

For illustrative purposes, we have selected some terms from the GEOLOGY domain, more specifically stratigraphical terms, taken from the DLPC. Stratigraphy is the branch of earth sciences that deals with stratified rocks. The OED defines it as 'the branch of geology concerned with the order and relative position of strata and their relationship to the geological timescale'. Saying 'the branch of' immediately conveys the idea of subordination to something. The OED definition allows us to say that `<Stratigraphy>` is a subordinate concept of `<Geology>`.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

The result of the application of the terminological methods gives rise to updated dictionary entries or senses for the DLP. Thus, the DLP has a double function: it will be both the corpus of analysis and the dictionary that will be improved with our methodological approach.

4. Terminological working methods for lexicographic work

Our methodological proposal has strictly lexicographic purposes and aims to employ terminological working methods to contribute to the treatment of specialised lexicographic content within general language dictionaries. The ultimate goal of our proposal is to offer strategies that can help lexicographers write accurate definitions. Meeting this need, we will address one of the most challenging tasks for any lexicographer – defining terms of subject fields they do not master.

The methodology we have followed assumes the completion of three essential stages: preparation, processing, and publishing. It is structured in ten phases to achieve the proposed objectives based on the theoretical assumptions mentioned before. Figure 2 presents the different phases that make up our methodology:

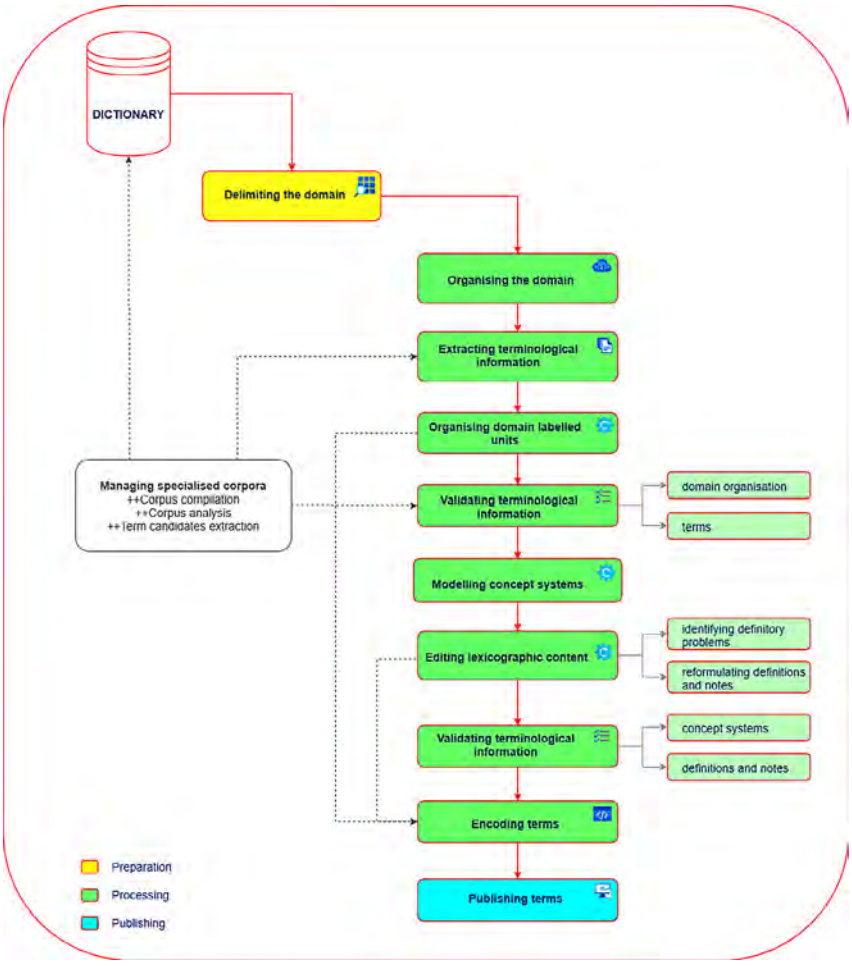


Fig. 2: Applying terminological methods when treating terms in general language dictionaries[2]

[2] For a detailed description of all the phases, see: Salgado (2021).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

We can identify some tasks that have a purely linguistic nature, such as the analysis of terms as designations of concepts, and other extralinguistic tasks that have a conceptual nature (ISO 704 2009), e. g., delimiting and organising domains, identifying concepts and concept relations, and modelling concepts systems.

Throughout this paper, we focus on three stages: 1) Organising the domain; 2) Modelling concept systems; and 3) Editing lexicographic content.

4.1 Organising the domain

Getting to know the domain and subsequently organising it are two requisite activities for a swift and systematic identification of the concepts, which will result in a better description of the set of terms.

As mentioned above, domain knowledge-building in dictionaries is achieved by resorting to a set of domain labels. We have analysed DLCP's labels and ended up suggesting the elimination of unnecessary or repetitive markings (Salgado/Costa/Tasovac 2021) and those distinctions that sometimes seem arbitrary because they are too narrow – both from a lexicographer's point of view and that of a regular user of the dictionary.

Concerning the labels related to the GEOLOGY domain, we have found four domain labels related to the broader concept of `<EarthSciences>` in the DLPC (CRYSTALLOGRAPHY, GEOLOGY, MINERALOGY, and PALAEONTOLOGY). In the absence of an explanation of the domain labelling system in the DLPC front matter, we consulted some specialised literature and found these same labels/descriptors in other existing classification systems (e. g., Dewey Decimal Classification (DDC);[3] Universal Decimal Classification (UDC);[4] EuroVoc;[5] UNESCO Thesaurus)[6] – see Table 1.

DLPC	METALABEL	Dewey Decimal Classification (DDC)	Universal Decimal Classification (UDC)	EuroVoc	UNESCO Thesaurus
cristalografia	**crystallography**	540 Chemistry & allied sciences/548 Crystallography	54 Chemistry. Crystallography. Mineralogy/548/549 Mineralogical sciences. Crystallography. Mineralogy		36 Science/2.20 Physical sciences/Cristallography
geologia	**geology**	550 Earth sciences/551 Geology, hydrology & meteorology	55 Earth sciences. Geological sciences / 551 General geology. Meteorology. Climatology. Historical geology. Stratigraphy. Paleogeography	36 SCIENCE/3606 natural and applied sciences/NT1 geology	2 Science/2.35 Earth sciences/Geology
mineralogia	**mineralogy**	540 Chemistry & allied sciences/549 Mineralogy	54 Chemistry. Crystallography. Mineralogy/548/549 Mineralogical sciences. Crystallography. Mineralogy	36 SCIENCE/3606 natural and applied sciences/NT1 geology/NT2 mineralogy	2 Science/2.35 Earth sciences/Mineralogy
paleontologia	**palaeontology**	560 Paleontology; paleozoology	56 Paleontology		2 Science/2.35 Earth sciences/Paleontology

Table 1: Comparison of DLPC domain labels and existing classification systems (Salgado/Costa/Tasovac 2021)

3 https://www.oclc.org/en/dewey.html.

4 http://www.udcsummary.info/php/index.php.

5 https://eur-lex.europa.eu/browse/eurovoc.html.

6 http://vocabularies.unesco.org/browser/thesaurus.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Taking the DDC or UDC from Table 1 as an example, the domains of CRYSTALLOGRAPHY and MINERALOGY are indexed in the class that covers CHEMISTRY. The fact that these domains are associated with CHEMISTRY, not with GEOLOGY, is acceptable since much of the subject actually falls into the CHEMISTRY domain; however, one cannot disregard that the subject is also directly related to GEOLOGY. Thus, interdisciplinarity is always a central point for domain organisation, and we have to take this into consideration when organising specialised knowledge in dictionaries. In Table 1, we have a column entitled *metalabel*, a tag that identifies the equivalent English designation of the corresponding domain. Using a metalabel will be beneficial for any work on aligning multiple dictionaries and studying them in parallel. This metalabel will also play an important role in the domain hierarchy that we will propose later on for the benefit of annotation.

Comparing the different classification systems has allowed us to offer a proposal to represent domains associated with EARTH SCIENCES in general language dictionaries applied to the DLP (Fig. 3), which the expert we consulted validated.

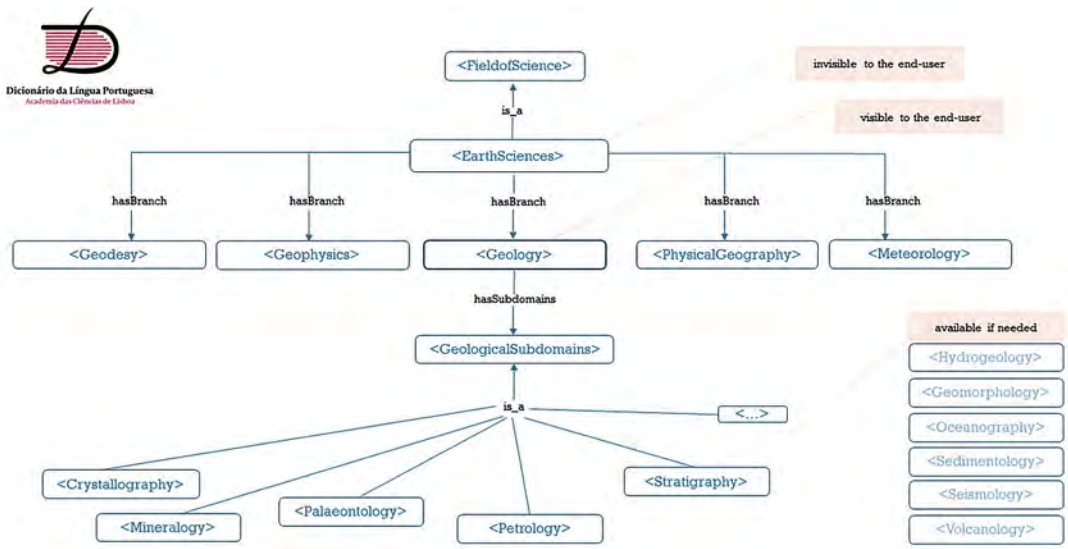


Fig. 3: Domain labels within the EARTH SCIENCES superdomain showing GEOLOGY as a domain and identifying its subdomains

This was the starting point to move from a non-hierarchical domain organisation to a hierarchical structure, which consequently increases the consistency of annotation and information retrieval. As Atkins/Rundell (2008) argued, instead of conceiving 'a totally flat non-hierarchical list of domains, it is more practicable to try to build a domain list with a certain hierarchical structure' (ibid., p. 184).

We have built hierarchical domain trees for lexicographic purposes. The hierarchy is as follows: superdomain, domain, subdomain (Salgado/Costa/Tasovac 2021). EARTH SCIENCES represent a superdomain followed by the domain GEOLOGY. In turn, GEOLOGY has various subdomains. Some domain labels will not be visible to the end-user since we consider labelling a lexicographic device for knowledge organisation. If the need to include other subdomains shall arise, they have already been foreseen. Our hierarchical domain trees can be made visible to give end-users the possibility of understanding the conceptual scope and how terms are interlinked, which is generally found isolated in general language dictionaries because they usually follow alphabetical order.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

To encode hierarchical domain labels, we used a customised version of TEI for lexicographic datasets – TEI Lex-0 (Tasovac et al. 2018) – employing the mechanism for the definition of taxonomies already available in `<teiHeader>`. This is possible in both plain TEI and TEI Lex-0 but has not been documented until now as a solution for representing usage labels. With this approach, domain labels are documented in `<encodingDesc>` (encoding description). The domains established in the taxonomy are declared in `<classDecl>` (classification declarations). This element is used to group the source of the domain's taxonomy used by the header or elsewhere in the document. First, the `<taxonomy>` element identifies the structured taxonomy. The categories are documented in the `<category>` element. Category elements are described, each defining a single category within the given taxonomy. Then, child categories are defined by the contents of a nested `<catDesc>` (category description) element, which contains the designation of the domain in the identified language. A single category may contain more than one `<catDesc>` child, and can be described in different languages (`xml:lang`). As a result of this thought process, we can establish a multilingual hierarchy for the EARTH SCIENCES superdomain (Fig. 4):

```
<encodingDesc>
    <classDecl>
        <taxonomy xml:id="domain">
            <category xml:id="domain.earth_sciences">
                <catDesc xml:lang="en">Earth Sciences</catDesc>
                <catDesc xml:lang="pt">Ciências da Terra</catDesc>
                <catDesc xml:lang="es">Ciencias de la Tierra</catDesc>
                <catDesc xml:lang="fr">sciences de la Terre </catDesc>
                <category xml:id="domain.earth_sciences.geology">
                    <catDesc xml:lang="en">Geology</catDesc>
                    <catDesc xml:lang="pt">Geologia</catDesc>
                    <catDesc xml:lang="es">Geología</catDesc>
                    <catDesc xml:lang="fr">Geologie</catDesc>
                    <category xml:id="domain.earth_sciences.geology.mineralogy">
                        <catDesc xml:lang="en">Mineralogy</catDesc>
                        <catDesc xml:lang="pt">Mineralogia</catDesc>
                        <catDesc xml:lang="es">Mineralogía</catDesc>
                        <catDesc xml:lang="fr">Mineralogie</catDesc>
                    </category>
                </category>
            </category>
        </taxonomy>
    </classDecl>
</encodingDesc>
```

Fig. 4: Hierarchical domain label for EARTH SCIENCES domain labels

Flat usage labels are usually encoded as text values of the `<usg>` element. For the sake of human readability, one could deploy the same strategy and explicitly add the domain label as the content of the `<usg>` element even when the full label taxonomy is maintained in the `<teiHeader>`. This would be particularly useful when the labels used in a given dictionary are not consistent.

Having organised the domains, we can start working with concepts.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

4.2 Modelling concept systems

Understanding concepts and the terms that denote them accurately depends on understanding the concept relations that interlink concepts in a concept system[7]. Our references were the concept relations and the graphic representations in the UML (Unified Modelling Language) notation proposed by the ISO 704 (2009) standard through concept diagrams[8].

We have identified hierarchical relations – generic[9] and partitive[10] – and associative[11] relations.

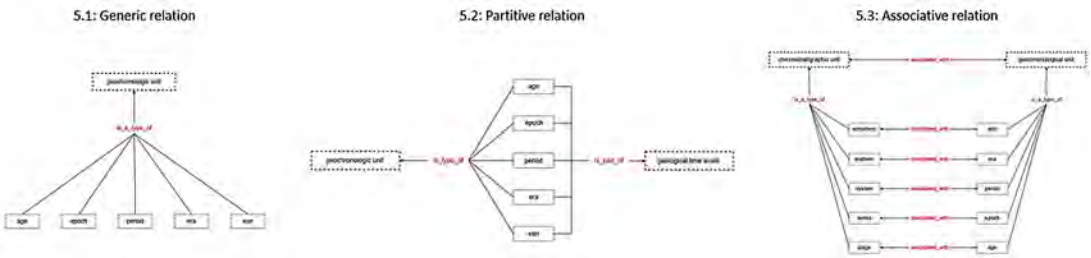


Fig. 5: Representation of conceptual relations using the concept of `<GeochronologicUnit>`

Here we have exemplified a generic concept relation (5.1) using the concept of `<GeochronologicUnit>` as a generic concept and `<Age>`, `<Epoch>`, `<Period>`, `<Era>`, and `<Eon>` as subordinate concepts. The specific concepts inherit a set of characteristics from their generic superordinate concept, i. e., the superordinate concept includes the subordinate concepts. The type of conceptual relation is made explicit using the marker `is_a_type_of`, which structures the generic/specific type relation. Regarding the semasiological approach, these markers also give us the possibility of detecting semantic relations, such as hypernym-hyponym relations. This exercise allowed us to detect that the superordinate concept `<GeochronologicUnit>` was not defined in the DLPC. Another argument required our attention: the subordination established between different concepts is not mirrored in the DLPC. These subordinate concepts constitute different entries in general language dictionaries, so one of the possible ways to represent the established semantic relations is to annotate them in TEI.

The primary means of conveying geological time information is through the Geological Time Scale and its units. Thus, all these units are part of the `<GeologicalTimeScale>` – a partitive relation (5.2). The conceptual relationship between the broader concept and its parts was made explicit through the conceptual marker `part_of`. Contrary to what was

7 A concept system is understood as a 'set of *concepts* structured in one or more related *domains* according to the *concept relations* among its concepts' (ISO 1087 2019, p. 6).

8 A concept diagram is a 'graphic representation of a *concept system*' (ISO 1087 2019, p. 7).

9 A generic relation exists between two concepts when the intension of the subordinate concept includes the intension of the superordinate concept plus at least one additional delimiting characteristic (ISO 704 2009, p. 9).

10 A partitive relation is said to exist when the superordinate concept represents a whole, while the subordinate concepts represent parts of that whole (ISO 704 2009, p. 13).

11 An associative relation exists when a thematic connection can be established between concepts (ISO 704 2009, p. 17).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

observed in generic relations, the principle of inheritance does not apply here, i. e., the concepts in a partitive relation do not inherit the characteristics of the superordinate concepts but do inherit their parts. The `<GeologicalTimeScale>` is a comprehensive concept, and all identified subordinate concepts – `<Age>`, `<Epoch>`, `<Period>`, `<Era>`, and `<Eon>` – represent parts of a whole, but they have distinctive characteristics concerning the related comprehensive concept.

To illustrate an associative concept relation (5.3), again we have used the concept of `<GeochronologicUnit>` in association with `<ChronostratigraphicUnit>`. We have a non-hierarchical relation: material–time, i. e., they have a semantic or pragmatic connection. If one wishes to allude to the time when these strata were deposited, then the concept of `<ChronostratigraphicUnit>` is replaced by that of `<GeochronologicUnit>`.

Once the conceptual relations are correctly identified, the lexicographer is able to start writing the definitions.

4.3 Editing lexicographic content

In this phase, senses are explained. For terminological purposes, a definition stabilises the relation between a concept and a term by the means of a linguistic expression. We distinguish the terminological definition (cf. De Bessé 1990; Rey 1995; Sager 2000; Temmerman 2000) from the lexicographic definition (Mel'čuk/Polguère 2018), which is generally suitable for general language dictionaries. Although both terminology and lexicography favour definition by intension, their purposes are different. The terminological definition attempts to state a concept designated by a term and characterise it in relation to other concepts within a concept system. In contrast, the lexicographic definition seeks to describe the signified meaning(s) of a lexical unit.

The terminological definition is related to the definition of the thing, as opposed to the lexicographic definition that relates to the usage of the word and is made by identifying the semantic features that characterise the meaning. The unit of meaning aimed at in the terminological definition is the concept. The difference between the terminological definition and the lexicographic definition, therefore, leads to different although not mutually exclusive approaches. In the context of general language dictionaries, the terminological definition has to be written for a non-expert audience.

ISO standards (ISO 704 2009; ISO 1087 2009) distinguish between intensional definition and extensional definition. The former consists of listing the immediate superordinate concept and delimiting the characteristics of the defined concept; the latter comprises listing its subordinate or partitive concepts. The definition by analysis or *genus-differentia* (Sager 1990) corresponds to ISO standards' intensional definition. Intensional definitions based on generic associations include the superordinate concept, followed by the delimiting characteristics within a concept system (e. g., `<Era>` among `<GeologicalTimeSpan>`). The superordinate concept's characteristics (that make up the intension) are assumed in the definition, which is the inheritance principle. Establishing conceptual relations facilitates the lexicographer's work and also enables the creation of a definitory model, e. g., `<GeochronologicUnit>` [superordinate concept] + `formed_during` [subordinate concepts].

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Existing definitions may have to be reformulated in cases where definitory problems may arise. On the other hand, the lexicographer can propose new definitions based on the previously established concept relations. For example, concerning the definition of "unidade geocronológica" [geochronologic unit], not included in the DLPC, we will suggest a definition considering the information retrieved from the following diagram (Fig. 6):

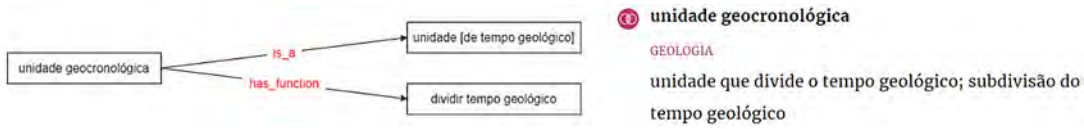


Fig. 6: Representation of the relation the conceptual markers `is_a` and `has_function` established from `<GeochronologicUnit>` and the unit "unidade geocronológica" defined in the DLP

As we can see in Figure 6, conceptual identifiers and linguistic markers may help lexicographers draft definitions. Focusing on the characteristics of a given concept is a fundamental step when defining it. The conceptual relation marker `is_a` establishes a hierarchical relation of subsumption. The conceptual marker `has_function` indicates the functionality of the unit. As we shall see later, we have assumed that these are instances of the so-called complex relationships (Sager 1990, pp. 34 f.), which are domain- and application- dependent. Thus, we have put forward the following definition for "unidade geocronológica" in the DLP: 'unidade que divide o tempo geológico; subdivisão do tempo geológico' [unit that divides geological time; geological time subdivision].

As lexicographers, we could not aim to work with all identified concepts. However, we consider it essential to analyse the relations among relevant concepts and organise them into concept systems, which will benefit the drafting of definitions. To illustrate this, Table 2 presents five different terms extracted from the DLPC and compares them with DLP's definitions that we have written after modelling the concept's microsystem. All of them define a type of `<GeochronologicUnit>`:

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

HEADWORD	DLPC (2001)	DLP (2021)
éon [eon]	*Geol.* longo período de tempo geológico que abarca duas ou mais eras	intervalo de tempo geológico (unidade geocronológica) durante o qual se formou um eonotema (unidade cronostratigráfica) Notas: 1) Na escala do tempo geológico, o éon é a categoria hierárquica mais elevada. 2) O éon integra várias eras.
era [era]	*Geol.* cada uma das grandes divisões do tempo geológico, cujos limites estão marcados por mudanças geológicas ou paleontológicas e que abrange vários períodos	intervalo de tempo geológico (unidade geocronológica) durante o qual se formou um eratema (unidade cronostratigráfica) Notas: 1) Na escala do tempo geológico, a era é hierarquicamente superior ao período e inferior ao éon. 2) A era integra vários períodos.
período [period]	—	intervalo de tempo geológico (unidade geocronológica) durante o qual se formou um sistema (unidade cronostratigráfica) Notas: 1) Na escala do tempo geológico, o período é hierarquicamente superior à época e inferior à era. 2) Na escala do tempo geológico, o período integra várias épocas.
época [epoch]	*Geol.* intervalo de tempo, nas divisões estratigráficas, que é relativo às formações de uma série ou conjunto de terrenos; subdivisão do período	intervalo de tempo geológico (unidade geocronológica) durante o qual se depositou uma série (unidade cronostratigráfica) Notas: 1) Na escala do tempo geológico, uma época é hierarquicamente superior à idade e inferior ao período. 2) Uma época integra várias idades.
idade [age]	—	intervalo de tempo geológico (unidade geocronológica) durante o qual se formou um andar (unidade cronostratigráfica) Notas: 1) A idade é a unidade básica da hierarquia do tempo geológico. 2) Quando necessário, a idade pode ser dividida em unidades geocronológicas de categoria inferior designadas por crono.

Table 2: Comparison of the definitions of "éon", "era", "período", "época", and "idade" in the DLPC (2001) and the DLP (2021)

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

For example, the lexicographic article "era" in the DLPC is 'Cada uma das grandes divisões do tempo geológico, cujos limites estão marcados por mudanças geológicas ou paleontológicas e que abrange vários períodos' [Each of the major divisions of geological time whose boundaries are marked by geological or palaeontological changes and which spans several periods]. This proposed definition lacks scientificity – it is too vague and even questionable as a formal statement. Comparing it to other Portuguese online dictionaries, we have found that "era" is defined in PRIBERAM (2021) as 'Divisão da escala de tempo geológico, superior ao período e inferior ao éon' [Division of geological time scale, higher than period and lower than eon] (PRIBERAM [emphasis added]). In INFOPÉDIA (2021), the lexicographic definition is 'unidade de divisão de tempo geológico, hierarquicamente inferior ao éon e superior ao período, definida por critérios paleontológicos e litológicos' [unit of geological time division, hierarchically lower than the eon and higher than the period, defined by palaeontological and lithological criteria] (INFOPÉDIA, 2021 [emphasis added]). On the contrary, and since we are modelling concept systems, we do not propose including this feature in the definition because the information given is not essential to define the given concept but may help to understand it. Instead, we recommend that additional information should be inserted as notes in the lexicographic article (in Table 2, see 'Notas' [Notes]).

Finally, if we observe the set of proposed definitions, the uniformity and systematisation in the treatment of terms are remarkable, highlighting the lack of systematisation in the previous edition. The analysis of the definitions according to the conceptual aspects is relevant in dictionaries even if the audience is not made up of experts.

5. Concluding remarks and further work

Following a terminologically-based approach improves the quality of the lexicographic product, both in terms of representation and organisation of knowledge and the description of terms themselves – the conceptual and linguistic dimensions. Combining these two different dimensions involves an iterative procedure. We should emphasise that we endorse the definition of the concept (ISO 1087 2019). In the DLP, we tested the creation of natural language definitions using concept systems. Focusing on the characteristics of a given concept is a fundamental step when defining it. We showed that conceptual identifiers and linguistic markers may help lexicographers draft definitions. As recommended by ISO 704 (2009), we conclude that intensional definitions are beneficial.

This work aims to facilitate the drafting of definitions, which, as we demonstrate, can be optimised and provided with greater scientific precision when we follow a terminological approach to the treatment of terms. The results obtained are immensely satisfactory, ensuring greater definition accuracy and quality. Instead of working a dictionary by classical alphabetical ordering (from A to Z), i. e., letter by letter, we found advantages in treating entries by sets of terms, first identifying the generic concept and describing its characteristics, and thus distinguishing it from other concepts.

By proposing hierarchical domain labels, we organise knowledge and establish higher and lower categories. The fact that we define a domain hierarchy does not mean that all proposed labels will be visible in the final product. This means that the lexicographer must structure the domains thoroughly and identify the terms according to the classification adopted. However, later on, the decision to make other domain categories visible to dictionary users must be weighed and considered taking into account the number of terms classi-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

fied under that label and also looking at the set of tags and their statistics in the realm of an established superdomain.

Given TEI Lex-0 has a non-standard nature (yet), it can be changed to accommodate relevant dictionary structures. We intend to demonstrate that the results obtained are helpful for computational lexical encoding and can serve the purpose of natural language processing. We have shown that the currently recommended TEI Lex-0 practice of representing domain labels as flat values is not robust enough to deal with more complex, hierarchical domain structures. The proposal that we present here for encoding hierarchical domain labels can be used in any dictionary, including multilingual ones. We recognise, however, that it is only a starting point for what we consider to be a joint effort to standardise domain labels, and that we have only dealt with two domains with a sampling of examples in each. In the future, we are also interested in exploring the results in the field of ontology.

We will continue to invest in an effective trans-disciplinary approach that combines theories and methods of terminology and lexicography, and even other disciplines, placing best practice standards at the core of our research. Unquestionably, terminology, with its interdisciplinary nature, is integral to knowledge conceptualisation and organisation, which justifies our approach.

References

ACL [Academia das Ciências de Lisboa] (2001): Dicionário da Língua Portuguesa Contemporânea, 2 vols. Casteleiro, J. M. (Coord.). Lisbon.

ACL [Academia das Ciências de Lisboa] (2021): Dicionário da Língua Portuguesa. Salgado, A. (Coord.). Lisboa. [New digital edition under revision.]

Atkins, B. T. S./Rundell, M. (2008): The Oxford guide to practical lexicography. New York.

Béjoint, H. (1988): Scientific and technical words in general dictionaries. In: International Journal of Lexicography 1 (4), pp. 354–368.

Boulanger, J.-C. (2001): L'aménagement des marques d'usage technolectales dans les dictionnaires généraux bilingues, dans Les dictionnaires de langue française. Dictionnaire d'apprentissage, dictionnaires spécialisés de la langue, dictionnaires de spécialité. In: Bibliothèque de l'Institut de linguistique française. Études de lexicologie, lexicographie et dictionnairique, n. 4, pp. 247–271.

Cabré, M. T. (1994): Terminologie et dictionnaires. In: Meta 39 (4), pp. 589–597. doi:10.7202/002182ar.

Cabré, M. T. (1999): Terminology: theory, methods and applications. Amsterdam. doi:10.1075/tlrp.1.

Costa, R. (2013): Terminology and specialised lexicography: two complementary domains. In: Lexicographica 29 (1), pp. 29–42. Doi:10.1515/lexi-2013-0004.

Costa, R. (2021): Terminology in the digital age: the ontological turn: Part 2. TOTh Training School 2021, 1–2 June 2021, France, Université Savoie Mont Blanc. Bourget du Lac.

De Bessé, B. (1990): La définition terminologique. In: Chaurand, J./Mazière, F. (ed.): Actes du Colloque la Définition, organisé par le CELEX (Centre d´études du Lexique) de l'Université Paris-Nord (1988). Paris, pp. 252–261.

Estopà, R. B. (1998): El léxico especializado en los diccionarios de lengua general: las marcas temáticas. In: Revista de la Sociedad Española de Linguística 28 (2), pp. 359–387.

Faber, P. (2009): The cognitive shift in terminology and specialized translation. In: MonTi: Monografías de Traducción e Interpretación 1, pp. 107–134. doi:10.6035/MonTI.2009.1.5.

Felber, H. (1987): Manuel de terminologie. Paris.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Gaudin, F. (2007): Socioterminologie: une approche sociolinguistique de la terminologie. Bruxelles.

Guerra Salas, L./Gómez Sánchez, M. (2005): El léxico especializado en los diccionarios monolingües de ELE. In: Castillo Carballo, M. A./Cruz Moya, O./García Platero, J. M./Mora Gutiérrez, J. P. (eds.): Actas del XV Congreso de Asele. Las gramáticas y los diccionarios en la enseñanza del español como segunda lengua: Deseo y realidad. Sevilla, pp. 427–434. https://cvc.cervantes.es/ensenanza/biblioteca_ele/asele/pdf/15/15_0425.pdf.

INFOPÉDIA (2021): Dicionário Infopédia da Língua Portuguesa. Porto Editora. https://www.infopedia.pt (last access: 24-03-2022).

ISO 1087 (2019): Terminology work – vocabulary – part 1: theory and application. Geneva.

ISO 704 (2009): Terminology work – Principles and methods. Geneva.

Mel'čuk, I./Polguère, A. (2018): Theory and practice of lexicographic definition. In: Journal of Cognitive Science 19 (4), pp. 417–470. doi:10.17791/jcs.2018.19.4.417.

Nomdedeu Rull, A. (2008): Hacia una reestructuración de la marca de 'deportes' en lexicografía. In: Azorín Fernández, D. et al. (Eds.), El diccionario como puente entre las lenguas y culturas del mundo. Actas del II Congreso Internacional de Lexicografía Hispánica. Alicante, pp. 764–770. https://dialnet.unirioja.es/servlet/articulo?codigo=5511595.

OED = Oxford English dictionary (2021). Oxford from https://www.oed.com (last access: 24-03-2022).

Paz Battaner, M. (1996): Terminología y diccionarios. In: Actes de la Jornada Panllatina de Terminologia. Barcelona, pp. 93–117.

PRIBERAM (2021): Dicionário Priberam da Língua Portuguesa. https://dicionario.priberam.org (last access: 24-03-2022).

Rey, A. (1985): La terminologie dans un dictionnaire général de la langue française: Le Grand Robert. In: TermNet News 14, pp. 5–7.

Rey, A. (1995): Essays on terminology. Amsterdam.

Roberts, R. P. (2004): Terms in general dictionaries. In: Bravo Gozalo, J. M. (ed.): A new spectrum of translation studies. Valladolid, pp. 121–140.

Roche, C. (2015): Ontological definition. In: Kockaert, H. J./Steurs, F. (eds.): Handbook of terminology. Vol. 1. Amsterdam, pp. 128–152.

Rondeau, G. (1984): Introduction à la terminologie. Montréal.

Sager, J. C. (1990): A practical course in terminology processing. Amsterdam.

Sager, J. C. (2000): Essays on definition. Amsterdam.

Salgado, A. (2021): Terminological methods in lexicography: conceptualising, Organising and encoding terms in general language dictionaries. Doctoral dissertation. Lisbon. https://run.unl.pt/handle/10362/137023.

Salgado, A./Costa, R./Tasovac, T. (2019): Improving the consistency of usage labelling in dictionaries with TEI Lex-0. In: Lexicography: Journal of ASIALEX 6 (2), pp. 133–156. doi:10.1007/s40607-019-00061-x.

Salgado, A./Costa, R./Tasovac, T. (2021): Mapping domain labels of dictionaries. In: Proceedings of XIX EURALEX International Congress: Lexicography for Inclusion. Alexandroupolis.

Simões, A./Almeida, J. J./Salgado, A. (2016): Building a dictionary using XML technology. In: Mernik, M./Leal, J. P./Oliveira, H. G. (Eds.), 5th Symposium on Languages, Applications and Technologies (SLATE'16) (14:1–14:8). Germany: Dagstuhl. Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik. doi:0.4230/OASIcs.SLATE.2016.0.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Tasovac, T./Romary, L./Bański, P./Bowers, J./Does, J./Depuydt, K./Erjavec, T./Geyken, A./Herold, A./Hildenbrandt, V./Khemakhem, M./Petrović, S./Salgado, A./Witt, A. (2018): TEI Lex-0: A baseline encoding for lexicographic data. Version 0.8.5. DARIAH Working Group on Lexical Resources. https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html.

Temmerman, R. (2000): Towards new ways of terminology description. The sociocognitive-approach. Amsterdam.

Tournier, J. (1992): Problèmes de terminologie en lexicologie anglaise et générale. Recherches en linguistique étrangère. XVI, pp. 215–226.

Wüster, E. (1979/1998): Introducción a la teoría general de la terminología y a la lexicografía terminológica. Barcelona: Institut Universitari De Lingüística Aplicada, Universitat Pompeu Fabra. [Einführung in die Allgemeine Terminologielehre und terminologische Lexikographie. Bonn 1979].

Contact information

Ana Salgado
NOVA CLUNL, Centro de Linguística da Universidade NOVA de Lisboa, Portugal/Academia das Ciências de Lisboa, Portugal
ana.salgado@fcsh.unl.pt

Rute Costa
NOVA CLUNL, Centro de Linguística da Universidade NOVA de Lisboa, Portugal
rute.costa@fcsh.unl.pt

Toma Tasovac
BCDH – Belgrade Center for Digital Humanities, Serbia
ttasovac@humanistika.org

Acknowledgements

This paper is supported by 1) the MORDigital – Digitalização do *Diccionario da Lingua Portugueza* de António de Morais Silva [PTDC/LLT-LIN/6841/2020] project financed by the Portuguese National Funding through the FCT – Fundação para a Ciência e Tecnologia and 2) the European Union's Horizon 2020 research and innovation programme under Grant agreement No 731015 (ELEXIS) (European Lexicographic Infrastructure).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Gilles-Maurice de Schryver

METALEXICOGRAPHY: AN EXISTENTIAL CRISIS

Abstract While there was arguably a need for multi-authored, multi-volume, metalexicographic handbooks three decades ago – when the field of metalexicography was still 'young' – it is a bit puzzling to make sense of the current output-flurry in this field. Is it simply a matter of 'every publisher trying to fill its shelves'? or is there really a need in the scientific community for more and (continuously) updated reference works? And once available, are such works also consulted? Which parts? By whom? How often? For what purposes? In this paper we look at an ongoing, real-world metalexicographic handbook project to answer these questions.

Keywords Metalexicography; major reference work; publishing model; download vs. citation patterns

1. The booming field of major metalexicographic reference works

When HSK 5.1 to 5.3 was published by Walter de Gruyter (eds. Hausmann et al. 1989–1991), Robert Ilson, then the editor of IJL, could claim "I use this enormous work all the time" (1997, p. 348); but apart from him, who was actually (and who has been) consulting this encyclopaedia (which runs over 3,400 pages)? Soon enough, the need for a fourth volume was felt, to fill in gaps that had been missed and to take account of the computer revolution; HSK 5.4 (with another 1,600 pages) eventually appeared, after a publication delay of five years, two decades later (eds. Gouws et al. 2013).

Fast-forward to the 2010s, and the community is at it again: Continuum published *e-Lexicography* (eds. Fuertes-Olivera/Bergenholtz 2011; 350 pages), and OUP *Electronic Lexicography* (eds. Granger/Paquot 2012; 500 pages). Then came *The Bloomsbury Companion to Lexicography* (ed. Jackson 2013; 450 pages), *The Oxford Handbook of Lexicography* (ed. Durkin (2016); 700 pages), *The Routledge Handbook of Lexicography* (ed. Fuertes-Olivera 2018; 850 pages), and most recently *The Cambridge Companion to English Dictionaries* (ed. Ogilvie (2020); 400 pages). On the day this contribution is being finalised,[1] the second edition of *The Bloomsbury Handbook of Lexicography* (ed. Jackson 2022; 510 pages) has just been released: alongside significantly updated and thoroughly revised chapters, it now includes a further six new chapters. Expected later in the year, is the *Cambridge Handbook of the Dictionary* (eds. Finegan/Adams 2022).

2. Case study: the *International Handbook of Modern Lexis and Lexicography* (IHMLL)

Crosscutting all these efforts is the *International Handbook of Modern Lexis and Lexicography* (IHMLL) (eds. Hanks/de Schryver 2014–2022), published by Springer as a so-called 'living reference work'. Initially planned to contain about 115 chapters (ranging from 10 to 80 pages each), chapters started appearing online in 2014 through 2017, after which the project stalled. Just 28 chapters had been finalised, or about a quarter (https://link.springer.com/

[1] March 24, 2022.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

referencework/10.1007/978-3-642-45369-4). There are quite a number of reasons for the standstill, but one of them, perhaps even the chief reason among them, was the result of an existential crisis: "What and who is all this really for?" The editors basically work for free, and all the authors contribute for free, while the publisher outsources the production to the cheapest corners of the world, necessitating endless revisions to multiple sets of proofs; yet once a chapter is finally 'ready' the publisher merely adds it to their ever-growing databases which contain literally hundreds of handbooks, totalling thousands of chapters, which they then bundle together with large numbers of digital books into 'eBook packages' for which academic libraries have to (and do!) pay eye-watering amounts, year after year.

1.1 Modern typesetting

There is clearly something very wrong with this modern business model; so wrong that it quite literally drove Patrick Hanks – otherwise no stranger to huge lexicographic projects – rather mad. The fact that such undertakings are often vanity projects, performed as an act of love for the field without proper remuneration is well known. But what is surely a new development is the extreme carelessness with which authors' texts are treated by typesetters. Never mind that many non-Latin letters (Greek, Cyrillic, even diacritics in languages such as Turkish or Vietnamese) come back garbled in the first set of proofs, an endless succession of 'project managers' at publishing houses is simply pushing buttons and handing down mindless instructions without anyone still actually checking anything for contents. Addenda 1 and 2 show two random examples: The first is an extract of page upon page in second (yes, second!) proofs filled with unreadable URLs; the second is an example of the mindboggling consequences of publishers employing people for the production process who clearly refuse to (or perhaps even cannot?) read. Proofs with utter gibberish are being presented with a straight face, and no amount of complaining seems to have any effect. One's project also moves from continent to continent, with new project managers being assigned every few months, so that the provocations never stop. It seems as if editors have to be grateful that big-name publishers are even willing to take on their projects, with the new normal now being that editors have to re-edit the work of the typesetters. Proofs are often so bad that editors 'hide' them from the authors, only sending through the second or third set of proofs for a final read.[2]

1.2 Publish as you go

Incomplete or not, this hasn't stopped the publisher to already include the IHMLL in its eBook packages, so that about a hundred libraries currently have access to the released quarter (https://worldcat.org/identities/lccn-n83-198535/).

Ironically, that released quarter led a life of its own over the past four years, as it has been available not only in the publisher's databases, but also in some of the authors' university repositories, as well as in academic social networks like ResearchGate (https://www.researchgate.net/) or Academia (https://www.academia.edu/). The four years that have elapsed since the stall thus provide us with natural access data.

[2] The problems with the way today's publishers behave, their new business models, and how they handle major reference works are not at all specific to metalexicography, alas.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

1.3 Bibliometrics

In an earlier bibliometric study (to appear in 2022) of the lexicographic journals *Dictionaries*, IJL, *Lexikos* and *Lexicographica*, it was found that 40% of the articles published in those journals are never ever cited. Two out of every five journal articles in lexicography could just as well never have been written, as no one ever refers to them. A cynic could even say, given that no colleague ever feels the need to refer to that material, that their only purpose was to populate one's own CV. In order to do better, we should all write less – 40% less. The problem, of course, is to know which two out of every five journal articles not to write.[3]

Moving from what are typically one-off contributions in lexicographic journals to chapters in major metalexicographic reference works, the question becomes: If two out of every five articles in lexicography don't attract even a single citation, what is the situation for chapters in handbooks on lexicography? Such handbooks are more akin to proper dictionaries, so whether or not a certain chapter ends up being cited (akin to whether or not a certain lemma ends up being looked up), is probably even more hit and miss. The preliminary assumption is therefore that the number of chapters never cited in a major handbook will be even higher than 40%.

Surprisingly, the data – for which, see Addendum 3 – reveal otherwise. According to Google Scholar (https://scholar.google.com/) just 9 of the 28 chapters have not (yet) been cited, or thus 32%. It is important to note, however – and when focusing on citations only, one tends to miss this important point – that zero citations doesn't mean that those chapters are never looked at. Indeed, the metrics on the publisher's website indicate that *all* chapters have been downloaded at least about a hundred times so far, up to nearly six hundred times for the most frequently downloaded ones. Hence, chapters which may not have attracted any citations so far, such as those on Icelandic, Yiddish, or etymology, have all been downloaded around a hundred times. We may thus assume that also those contributions were not written in vain after all. The three most downloaded chapters are 'Dictionaries as aids for language learning' (574 x), 'Historical principles vs. synchronic approaches' (566 x), and 'The lexicography of Khmer' (468 x). The three most cited chapters are 'Dictionaries and their users' (30 x), 'Dictionaries and crowdsourcing, wikis and user-generated content' (19 x), and 'Bilingual lexicography: translation dictionaries' (19 x).

The correct way to look at the values, though, is to take the actual number of days each chapter has been online into account, which is why Addendum 3 also includes columns with the download and citation values expressed 'per year', meaning 'per 365 days'. Note that the rows of Addendum 3 are ordered, first, from highest to lowest 'citations/year', and second, for those without any citations, from highest to lowest 'downloads/year'.

1.4 Downloads vs. citations

Modern dictionaries are based on corpus data, and an interesting metalexicographical discussion of the past 15 years concerns the question as to whether or not there is a correlation between corpus frequency and look-up frequency. While initial results using Bantu data

[3] While this figure may be shocking to many, it is actually pretty respectable for the humanities, where up to 82% of published articles may go uncited (see Remler 2014, citing evidence from Larivière et al. 2009). For more on the power-law behaviour in the distribution of citations, see for instance Brzezinski (2015).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

suggested that there is no useful relationship (de Schryver et al. 2006), this was challenged using a better methodology with Indo-European data (Koplenig/Meyer/Müller-Spitzer 2014; Trap-Jensen/Lorentzen/Sørensen 2014; Müller-Spitzer/Wolfer/Koplenig 2015), up-on which two of the teams joined hands, to indeed conclude that there is a positive correlation after all, one which they now even claim is universal (de Schryver/Wolfer/Lew 2019).

A parallel question may also be asked of the IHMLL, namely: *To what extent do download numbers predict citation patterns in metalexicography?* At face value, one would assume that more downloads lead to more citations, and that fewer downloads will inevitably result in fewer citations. Researchers have for instance shown this to be the case in the field of analytical chemistry (Jahandideh/Abdolmaleki/Barzegari Asadabadi2007), but our (admittedly very limited) data for the field of metalexicography do not seem to corroborate that. Figure 1 brings the two values together for the first 28 chapters of the IHMLL: on the *x*-axis the number of downloads/year, on the *y*-axis the number of citations/year. The top three downloaded chapters are all found in the bottom-right of the graph, where they barely attracted citations. Conversely, it is those chapters with medium numbers of downloads/year, i. e. those in the top half of the graph, which resulted in the most citations/year. The 9 chapters with no citations of course pull the trendline down, revealing one additional outlier, viz. the chapter on 'Natural Semantic Metalanguage and lexicography'.

Overall, and over the past few years, each chapter was downloaded an average of 45 times a year, and cited an average of 0.74 times a year. And this while the project has been basically dormant, with no advertisement of it whatsoever, nor even a word about its very existence.[4]

[4] The sum of the downloads for the 28 chapters in Addendum 3 is 7,487, while the Springer website claims "9.2k" for this metric. Following an enquiry, the Springer 'project coordinator' for the IHMLL at the time confirmed that "due to a temporary bug" the overall metric shown online is about 2,000 downloads too high (personal communication, e-mail 16/02/2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

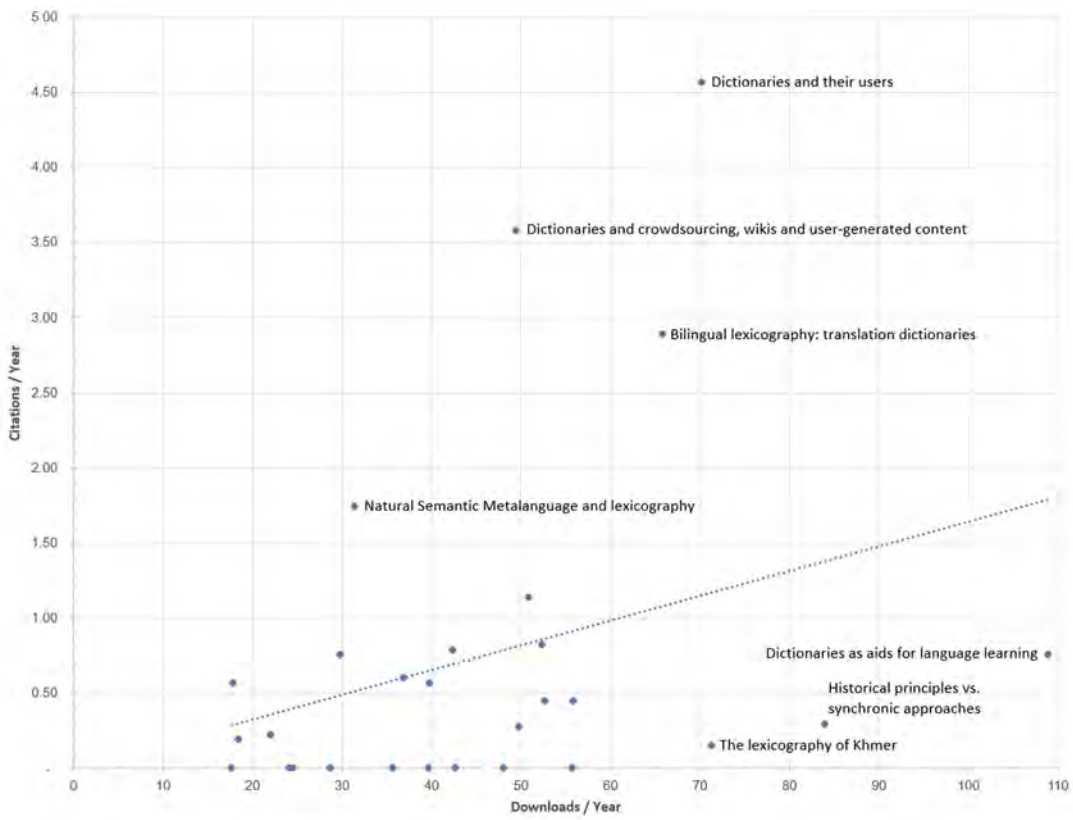


Fig. 1: Downloads/year vs. citations/year for the IHMLL

The positioning of the dots in the graph of Figure 1 is intriguing, and begs for a comparison with more general trends in metalexicography. For earlier work in this field, see de Schryver (2009a, 2009b, 2012a, 2012b, 2019) and Lew/de Schryver (2014). The quickest 'single-shot, one-stop' place to assess Figure 1 is probably to allow Google Scholar (GS) to do the heavy lifting. Bringing the world's lexicographers with a GS profile (and the label 'lexicography') together, as done here https://scholar.google.com/citations?view_op=search_authors&hl=en&mauthors=label:lexicography, immediately reveals that most citations in metalexicography are collected by those colleagues involved in the interface between language teaching and lexicography (cf. Batia Laufer and Sylviane Granger in positions 1 and 3). The users of the IHMLL follow this pattern in terms of top download (cf. 'Dictionaries as aids for language learning') but, as seen above, do not follow this up with citations. Similarly, 'Historical principles vs. synchronic approaches' is an old favourite in traditional metalexicography (cf. Patrick Hanks in position 2), again not followed up in terms of citations. Conversely, the new metalexicographic topics which focus on digital users, crowdsourcing, etc. are only found on pages two and three (each page lists 10 colleagues) at GS (cf. e. g. Michael Rundell and Robert Lew). What this suggests is that the user of the IHMLL consults this handbook to move the modern aspects of our field forward, engaging with and citing the material, while the older topics, while consulted (proxy: 'downloaded') more often, are mainly browsed out of interest.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

3. Outlook

In conclusion we can thus state that handbooks such as the IHMLL do have an audience and a purpose after all. Work on the IHMLL has therefore restarted in earnest: Over the past half year a further 22 chapters were put into production, which will bring the total to 50. By the time of the Euralex 2022 congress, we expect to be closer to the end. Here's hoping.

References

Brzezinski, M. (2015): Power laws in citation distributions: evidence from scopus. In: Scientometrics 103 (1), pp. 213–228.

de Schryver, G.-M. (2009a): Bibliometrics in lexicography. In: International Journal of Lexicography 22 (4), pp. 423–465.

de Schryver, G.-M. (2009b): Lexikos at eighteen: an analysis. In: Lexikos 19, pp. 372–403.

de Schryver, G.-M. (2012a): Lexicography in the crystal ball: facts, trends and outlook. In: Fjeld, Ruth V./Torjusen, Julie M. (eds.): Proceedings of the 15th EURALEX International Congress, 7–11 August, 2012, Oslo. Oslo, pp. 93–163.

de Schryver, G.-M. (2012b): Trends in twenty-five years of academic lexicography. In: International Journal of Lexicography 25 (4), pp. 464–506 + 42 pages of supplementary material online.

de Schryver, G.-M. (2019): Past, present and future in Asian lexicography. Paper presented at the 13th International Conference of the Asian Association for Lexicography, ASIALEX 2019, Istanbul, Turkey, 19–21 June 2019.

de Schryver, G.-M./Wolfer, S./Lew, R. (2019): The relationship between dictionary look-up frequency and corpus frequency revisited: a log-file analysis of a decade of user interaction with a Swahili-English dictionary. In: GEMA Online® Journal of Language Studies 19 (4), pp. 1–27 + supplementary material online.

de Schryver, G.-M./Joffe, D./Joffe, P./Hillewaert, S. (2006): Do dictionary users really look up frequent words? – On the overestimation of the value of corpus-based lexicography. In: Lexikos 16, pp. 67–83.

Durkin, P. (ed.) (2016): The Oxford handbook of lexicography. New York.

Finegan, E./Adams, M. (eds.) (2022): The Cambridge handbook of the dictionary. Cambridge.

Fuertes-Olivera, P. A. (ed.) (2018): The Routledge handbook of lexicography. London.

Fuertes-Olivera, P. A./Bergenholtz, H. (eds.) (2011): e-Lexicography: the internet, digital initiatives and lexicography. London.

Gouws, R. H./Heid, U./Schweickard, W./Wiegand, H. E. (eds.) (2013): Dictionaries. An international encyclopedia of lexicography. Supplementary volume: Recent developments with focus on electronic and computational lexicography. (= Handbooks of Linguistics and Communication Science (HSK) 5.4). Berlin.

Granger, S./Paquot, M. (eds.) (2012): Electronic lexicography. Oxford.

Hanks, P./de Schryver, G.-M. (eds.) (2014–2022): International handbook of modern lexis and lexicography. Berlin.

Hausmann, F. J./Reichmann, O./Wiegand, H. E./Zgusta, L. (eds.) (1989–1991): Wörterbücher/Dictionaries/Dictionnaires. Ein internationales Handbuch zur Lexikographie/An international encyclopedia of lexicography/Encyclopédie internationale de lexicographie. (= Handbücher zur Sprach- und Kommunikationswissenschaft (HSK) 5.1–5.3). Berlin.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Ilson, R. F. (1997): Review: Wörterbücher/Dictionaries/Dictionnaires: An international encyclopedia of lexicography. Edited by Franz Josef Hausmann, Oskar Reichmann, Herbert Ernst Wiegand, Ladislav Zgusta. Berlin/New York. Vol. 1, 1989; Vol. 2, 1990; Vol. 3, 1991. pp. LII+3355. In: International Journal of Lexicography 10 (4), pp. 348–357.

Jackson, H. (ed.) (2013): The Bloomsbury companion to lexicography. London.

Jackson, H. (ed.) (2022): The Bloomsbury handbook of lexicography (= Bloomsbury Handbooks). London.

Jahandideh, S./Abdolmaleki, P./Barzegari Asadabadi, E. (2007): Prediction of future citations of a research paper from number of its internet downloads. In: Medical Hypotheses 69 (2), pp. 458–459.

Koplenig, A./Meyer, P./Müller-Spitzer, C.n (2014): Dictionary users do look up frequent words. A log file analysis. In: Müller-Spitzer, C. (ed.). Using Online Dictionaries (= Lexicographica. Series Maior 145). Berlin, pp. 229–249.

Larivière, V./Gingras, Y./Archambault, É. (2009): The decline in the concentration of citations, 1900–2007. In: Journal of the American Society for Information Science and Technology 60 (4), pp. 858–862.

Lew, R./de Schryver, G.-M. (2014): Dictionary users in the digital revolution. In: International Journal of Lexicography 27 (4), pp. 341–359.

Müller-Spitzer, C./Wolfer, S./Koplenig, A. (2015): Observing online dictionary users: studies using Wiktionary log files. In: International Journal of Lexicography 28 (1), pp. 1–26.

Ogilvie, S. (ed.) (2020): The Cambridge companion to English dictionaries. Cambridge.

Remler, D. (2014): Are 90% of academic papers really never cited? Searching citations about academic citations reveals the good, the bad and the ugly. https://dahliaremler.com/2014/04/09/are-90-of-academic-papers-really-never-cited-searching-citations-about-academic-citations-reveals-the-good-the-bad-and-the-ugly/.

Trap-Jensen, L./Lorentzen, H./Sørensen, N. H. (2014): An odd couple – corpus frequency and look-up frequency: what relationship? In: Slovenščina 2.0 2 (2), pp. 94–113.

Contact information

Gilles-Maurice de Schryver
BantUGent – UGent Centre for Bantu Studies, Ghent University
&
Department of African Languages, University of Pretoria
gillesmaurice.deschryver@UGent.be

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Addendum 1: Example of unreadable second proofs

the Greek language of the seventeenth to nineteenth centuries. There are some valuable resources for that period such as Johannes Meursius' *Glossarium Graeco-Barbarum* (1614) (more at https://en.wikipedia.org/wiki/Johannes_Meursius), Alessio da Somavera's *Tesoro della lingua greca-volgare ed italiana* (1709) (which can be accessed at http://webcache.googleusercontent.com/search?q=cache:oglfkpQdkDsJ:anemi.lib.uoc.gr/metadata/2/4/b/metadata-01-0001129.tkl+&cd=1&hl=el&ct=clnk&gl=gr), and Skarlatos Vyzantios' Dictionary (1835) (which can be accessed at http://anemi.lib.uoc.gr/search/?dtab=m&search_type=simple&search_help=&display_mode=overview&wf_step=init&show_hidden=0&number=10&keep_number=10&cclterm1=&cclterm2=&cclterm3=&cclterm4=&cclterm5=&cclterm6=&cclterm7=&cclterm8=&cclfield1=&cclfield2=&cclfield3=&cclfield4=&cclfield5=&cclfield6=&cclfield7=&cclfield8=&cclop1=&cclop2=&cclop3=&cclop4=&cclop5=&cclop6=&cclop7=&isp=&display_help=0&offset=11&search_coll%5bmetadata%5d=1&&stored_cclquery=creator%3D%28%CE%92%CF%85%CE%B6%CE%AC%CE%BD%CF%84%CE%B9%CE%BF%CF%82%2C+%CE%A3%CE%BA%CE%B1%CF%81%CE%BB%CE%AC%CF%84%CE%BF%CF%82+%CE%94.%2C%29&skin=&rss=0&show_form=&export_method=none&ioffset=1&display_mode=detail&ioffset=1&offset=13&number=1&keep_number=10&old_offset=11&search_help=detail), et al. Stefanos Koumanoudis attempts to gather in his "Collection" (1900) (which can be accessed at http://anemi.lib.uoc.gr/search/?dtab=m&search_type=simple&search_help=&display_mode=overview&wf_step=init&show_hidden=0&number=10&keep_number=10&cclterm1=&cclterm2=&cclterm3=&cclterm4=&cclterm5=&cclterm6=&cclterm7=&cclterm8=&cclfield1=&cclfield2=&cclfield3=&cclfield4=&cclfield5=&cclfield6=&cclfield7=&cclfield8=&cclop1=&cclop2=&cclop3=&cclop4=&cclop5=&cclop6=&cclop7=&isp=&display_help=0&offset=11&search_coll[metadata]=1&&stored_cclquery=creator=(Κουμανο

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Addendum 2: Example of typesetters sleeping at the wheel

Author input:

211 *Perkamusan Indonesia* (1976) lists no less than 16 post-independence dictionaries
212 intended for an Indonesian readership, and many more have appeared since. Among
213 the most exhaustive is a 756-page Javanese-Indonesian dictionary by
214 Prawiroatmodjo (1957).

user

layout?

Typesetter action:

Since Indonesia's independence, Javanese remained one of the best documented languages, second only to Malay/Indonesian. The aforementioned *Bibliografi Perkamusan Indonesia* (1976) lists no less than 16 post-independence dictionaries intended for an Indonesian readership, and many more have appeared since. Among layout? Javanese-Indonesian dictionary by Prawiroatmodjo (1957).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Addendum 3: Download and citation data for the IHMLL so far (28 chapters, up to 24/03/2022)

Chapter title	Author(s)	Online since	Days online	Down-loads up to 24/03/2022	Down-loads / year	Citations up to 24/03/2022	Citati-ons / year
Dictionaries and their users	Robert Lew	03/09/2015	2,400	462	70.26	30	4.56
Dictionaries and crowdsourcing, wikis and user-generated content	Michael Rundell	07/12/2016	1,939	263	49.51	19	3.58
Bilingual lexicography: translation dictio-naries	Arleta Adamska-Sałaciak	03/09/2015	2,400	433	65.85	19	2.89
Natural Semantic Metalanguage and lexicography	Cliff Goddard	30/08/2017	1,673	144	31.42	8	1.75
Construction grammar and lexicography	Bill Croft/Logan Sutton	21/12/2016	1,925	268	50.82	6	1.14
Term banks	Thierry Fontenelle/Dieter Rummel	06/12/2014	2,671	383	52.34	6	0.82
Figurative language and lexicography	Alice Deignan	17/11/2015	2,325	270	42.39	5	0.78
Dictionaries as aids for language learning	Alex Boulton/Sylvie De Cock	22/12/2016	1,924	574	108.89	4	0.76
The lexicography of Sephardic Judaism	David M. Bunis	23/12/2016	1,923	157	29.80	4	0.76
The lexicography of German	Annette Klosa	06/04/2017	1,819	184	36.92	3	0.60
Sign language lexicography	Rachel McKee/Mireille Vale	22/12/2016	1,924	210	39.84	3	0.57
The lexicography of Portuguese	Ana Frankenberg-Garcia	20/12/2016	1,926	94	17.81	3	0.57
The lexicography of Indonesian/Malay	Deny Arnos Kwary/ Nor Hashimah Jalaluddin	14/07/2015	2,451	375	55.84	3	0.45
The lexicography of indigenous languages in Australia and the Pacific	Nick Thieberger	01/08/2015	2,433	351	52.66	3	0.45

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Chapter title	Author(s)	Online since	Days online	Down-loads up to 24/03/2022	Down-loads / year	Citations up to 24/03/2022	Citati-ons / year
Historical principles vs. synchronic approaches	Judy Pearsall	06/07/2015	2,459	566	84.01	2	0.30
The lexicography of indigenous languages in South America	Wolf Dietrich	06/12/2014	2,671	364	49.74	2	0.27
Norms and exploitations in lexicography	Sara Može	13/09/2017	1,659	100	22.00	1	0.22
The lexicography of Scots	Susan Rennie	11/01/2017	1,904	96	18.40	1	0.19
The lexicography of Khmer	Robert K. Headley	08/09/2015	2,395	468	71.32	1	0.15
The lexicography of minority languages in Southeast Asia	David Bradley	03/07/2015	2,462	376	55.74	0	0
The lexicography of Persian (Farsi, Tajiki, and Dari)	Corey Andrew Miller	25/11/2017	1,586	209	48.10	0	0
The lexicography of Hebrew	Tsvi Sadan	10/01/2017	1,905	223	42.73	0	0
The lexicon of the male sex worker	Welby Ings	13/11/2015	2,329	253	39.65	0	0
The lexicography of Norwegian	Oddrun Grønvik	21/11/2016	1,955	191	35.66	0	0
The lexicography of Esperanto	Federico Gobbo	03/12/2016	1,943	153	28.74	0	0
Etymology in dictionaries	Anatoly Liberman	06/09/2017	1,666	112	24.54	0	0
The lexicography of Yiddish	Wolf Moskovich	21/06/2017	1,743	115	24.08	0	0
The lexicography of Icelandic	Þórdís Úlfarsdóttir/Kristín Bjarnadóttir	20/12/2016	1,926	93	17.62	0	0
		AVERAGES	**2,083**	**267**	**45.24**	**4**	**0.74**

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Corpora in Lexicography

EURALEX XX 22
XX EURALEX INTERNATIONAL CONGRESS
DICTIONARIES AND SOCIETY
12-16 July 2022, Mannheim, Germany

Nils Diewald/Marc Kupietz/Harald Lüngen

TOKENIZING ON SCALE

Preprocessing large text corpora on the lexical and sentence level

Abstract When comparing different tools in the field of natural language processing (NLP), the quality of their results usually has first priority. This is also true for tokenization. In the context of large and diverse corpora for linguistic research purposes, however, other criteria also play a role – not least sufficient speed to process the data in an acceptable amount of time. In this paper we evaluate several state-of-the-art tokenization tools for German – including our own – with regard to theses criteria. We conclude that while not all tools are applicable in this setting, no compromises regarding quality need to be made.

Keywords Corpora; tokenization; software

1. Introduction

Tokenization, that is, the segmentation of texts into lexical units, is a fundamental preprocessing step for lexicographic work with corpus linguistic resources. Although tokenization is one of the simpler tasks in these processing steps, it is often critical because early errors can affect all analyses and procedures based on it. Accordingly, high accuracy in tokenization is usually the outstanding criterion in the evaluation of tools used for this purpose. Depending on the area of application, other criteria may also play an important role in the evaluation of tokenization tools. The evaluation presented here is based on a scenario in research data preparation, more precisely: the tokenization of DeReKo, the German Reference Corpus. With currently more than 50 billion running words, it is a very large and constantly growing linguistic data resource, for which not only high accuracy is relevant for tokenization, but also a speed that allows the resource to be processed in an acceptable time. Other criteria in this scenario is the extensibility of the language model for new linguistic phenomena and the adaptability for new or different corpora. Also important is the permanent maintainability of the tools and the reproducibility of its results for research. In this article, we evaluate different tools for tokenizing German text data with a special focus on the scenario outlined here. Furthermore, we include sentence segmentation in our consideration, since it is often performed in the same processing step. We include two of our own implementations in the evaluation and compare them with several off-the-shelf tools that we consider state-of-the-art.

2. Tokenization

In corpus technology, tokens represent basic lexical units which are indexed and can be addressed in search queries to the corpus. In fact it is sometimes difficult to query for characters that are delimiters in tokenization. The unsegmented characters of a corpus text are considered its primary data, and its tokenization serves as a basic layer for searching and for higher-level analyses provided as annotations, such as part-of-speech tagging, named entity recognition, syntactic parsing, or anaphora resolution. The tokenization scheme is therefore crucial for the analyses that can be expressed and represented on the higher levels. Further-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

more, tokens are the basis for the calculation of the corpus size and statistical measures based on the corpus size and/or token frequencies.

For languages with alphabetical writing systems that use spaces to mark word boundaries (such as German), a simple tokenization algorithm consists of using these spaces and punctuation symbols of a text as delimiters and to consider the resulting strings between them and the punctuation symbols themselves as tokens. This method leads to meaningful tokens in the majority of cases already (s. sec. 4.3).

The bulk of the difficulties with tokenization arises from the potential ambiguity of the space and punctuation characters, as in certain cases they should not be considered token delimiters but instead parts of tokens. Certain multi-word expressions, such as "ad hoc", "bus driver" or "heart attack" in English should under a syntactic perspective be analysed as single lexical tokens that just happen to contain the space character. As for punctuation characters, there are several cases of when the dot '.' does not represent the full stop terminating a sentence e. g. in abbreviations ("etc."; "Fa.", short for *Firma*, 'company'; "bzw.", short for *beziehungsweise*, 'respectively'), in ordinary numbers and enumerations ("1.", "a.)", "B."), as part of email addresses or URLs, and in several other cases (cf. Proisl/Uhrig 2016). Similarly, other punctuation symbols are ambiguous between delimiters and other uses, for instance the hyphen in "Hamburg-München" (naming a distance, leading to three separate tokens) vs. in "Reich-Ranicki" (a surname, I. e. one token), or in the brand name "Yahoo!", the "!" should not be separated. Non-alphabetic characters other than punctuation can also be ambiguous between representing a delimiter vs. a regular character that is part of a token, e. g. in an arithmetic expression like "5+3", the "+" is a delimiter, but in "C++", it should not be separated, also cf. The asterisks in an action word like "*grins*" (from *grinsen*, 'to grin'; separate tokens according to the Tokenization Guidelines of the EmpiriST Shared Task; Beißwenger et al. 2015) vs. in a form like "Lehrer*innen" (gender neutral form for 'teachers'; one token). In turn, there are also cases when strings without delimiters actually contain separate tokens, mostly when spaces are intentionally or unintentionally omitted as in "er-hat" for *er hat* ('he has').

In our recent corpora, we have identified five types of phenomena, mostly connected with the internet as a media for the distribution of texts, which pose relatively new challenges for tokenization by introducing additional ambiguities, or significantly increasing the frequency of occurrence of certain known ambiguities.

1) The proliferation of unedited text i. e. texts that contain sloppy or creative uses of spelling, which can also affect the tokenization, as in contracted forms based on the spoken language e. g. "haste" (→ *hast du*; 'do you have'), "hastes" (→ *hast du es*, 'do/here you have it'), the omission of spaces ("Hänselundgretel", *Hänsel und Gretel*, 'Hansel and Gretel'), the insertion of extra spaces ("Auto Bahn", *Autobahn*, german highway system), or an iterative use of punctuation symbols ("Drogen!!!!", 'Drugs!!!!') (cf. Bartz/Beißwenger/Storrer 2013).

2) The proliferation of computer-mediated communication (CMC) idioms used in social media which besides displaying features of the spoken language as mentioned above, contain new specific uses of punctuational and non-alphabetical characters as in emoticons such as ":-)", addressing terms ("@heiner"), or action words and phrases ("*lach*", from *lachen*, 'to laugh'; "*auf die Nägel blas*", 'to blow on the nails') (cf. Bartz/Beißwenger/Storrer 2013).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

3) New, token-internal uses of delimiters in gender-conformant spellings as in German forms like "Lehrer(-in)", "Lehrer:in", "Lehrer/-innen", "Lehrer/innen", "Lehrer_innen", "Lehrer*innen" (gender neutral forms for 'teachers'), "diese*r" (gender neutral form for 'this/these/those'), "Frau*".

4) Hypertextual features such as hashtags, mentions, filenames, email addresses, URLs, but also XML or HTML markup, Markdown, or WikiCreole source code (see Jurish/Würzner 2013).

5) The huge quantities of text that the internet offers pose severe processing challenges for tokenization in terms of time and space.

Note that different NLP applications or a different focus of linguistic description may require different tokenization strategies. For instance, when the focus is on syntax/parsing, compounds are analysed as one token, whereas for semantic relation or information extraction, it might be relevant to even tokenize the parts of compounds that are spelt as one word (such as "Abgasrückführung", 'exhaust gas recycling'). In fact, different NLP tools often disagree in their tokenization (s. sec. 4.3), and their tokenization strategies cannot independently be considered as correct or incorrect. The interpretation of tokens we adopt is closer to a lexicographic reading and might deviate from the preprocessing in some machine-learning (ML) workflows that are based on dictionary queries and tokenize other units to handle out-of-vocabulary situations. We also do not consider tokenizations into sub-lexical units such as morphemes.

The task of sentence segmentation is closely related to tokenization due to the central role of punctuation symbols and their ambiguity between being a part of a token or terminating a sentence. In fact, tokenization and sentence segmentation are often applied in the same processing step. Besides the ambiguity of punctuation, sentences or sentence-like units might not be delimited by a full stop at all, as for example regularly in the case of headings, but also in instances of sloppy writing in CMC. Sentences represent the basic scope of later syntactic analyses, and a faulty sentence segmentation renders automatic syntactic parsing largely invalid. In corpus technology, the sentence is also the default domain of a query, i. e. when querying for and analysing expressions with multiple parts (as in "ADJ N"), it is crucial that no sentence boundary lies between them. Moreover, various licence-related restrictions refer to the unit *sentence*. Consequently, this unit is of great, not only linguistic, importance in the DeReKo scenario.

3. Large Scale Scenario

Our main use case for corpus tokenization represents the preprocessing of the German reference corpus DeReKo (Kupietz et al. 2018). DeReKo has been compiled at the Leibniz Institute for the German Language since 1964 and currently comprises more than 50 billion running words with an annual increase of more than 2 billion. It is used for a broad range of linguistic research on written contemporary German and for this purpose is completely tokenized, and morphologically and syntactically annotated multiple times. For researchers, access is limited (primarily for licensing reasons) to search engines (COSMAS II and KorAP) and further analysis tools. In addition, various derived analysis data such as frequency-based wordlists are offered. These different forms of usage also determine the pragmatic token definition on which DeReKo is based, which tries to be both "linguistically significant and methodologically useful" (Webster/Kit 1992, p. 1106), although compromises must be made

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

for both aspects. This is particularly important to consider with respect to "word" tokens, which must follow a lexicographic definition. Idioms and fixed expressions, which consist of several, possibly even discontinuous units, can be useful in a lexicological sense, but detrimental for search engine use, which is why DeReKo does not take them into account. Consequently, we assume, on the one hand, tokens to be the "minimal unit of investigation" (Chiarcos/Ritz/Stede 2009, p. 35), and, on the other hand, units that are meaningful for representations in syntactic contexts.

With respect to the data, both in terms of size and diversity, there are further pragmatic requirements for tokenization. In our evaluation, we emphasize high processing **speed** with large data volumes and limited resources. In the case of DeReKo, this is important with respect to the constant acquisition of data, but is especially necessary when a complete re-tokenization of the entire corpus is required, for example, to correct systematic errors while maintaining model consistency. The limited technical resources also mean, in our scenario, that the procedures should run on commodity hardware – accordingly, we do not consider possible performance increases through special hardware (such as GPU support) in this analysis.

Furthermore, high **quality** is of obvious importance, since tokenization is the basis of further linguistic analysis steps, and, as Moreau/Vogel (2018, p. 1120) correctly note: "Tokenization errors can be costly performance-wise, as these errors may propagate through the whole processing chain."

A controllable **extensibility** of tokenization is of great importance especially with respect to new phenomena and new, special corpora. In some cases, prior linguistic knowledge can be used for extension, but also specially prepared training corpora. What plays only a minor role for our application scenario, but is primary in many other scenarios, is the **adaptability** to other languages and corpora. Even though we have adapted our tools for other languages, the focus is on German language data.

With regard to the use of processed data for research purposes, other important aspects are the **maintainability** of the solutions and the **reproducibility** of the results. For maintainability, it is necessary that the solutions are available as open source, which is why we restrict the evaluation exclusively to this. Applicability to commodity hardware has already been mentioned but is also central in terms of maintainability. In terms of reproducibility, the results, but also the errors, should be consistent and comparable across genre and domain boundaries. Along with this, a desirable advantage is the reduction of the tokenization step to as few tools as possible. Different tokenizers for heterogeneous data, adapted to different corpus types, genres, or domains, would not only complicate the traceability and thus reproducibility of any tokenization errors, they would also pose a major challenge in the handling and maintainability of the full preparation pipeline.

Not to be considered in our scenario are preprocessing steps that may be necessary with respect to the specific origin of the data. Since the source data for DeReKo usually come in XML or other preprint format, there is no need to perform print-specific preprocessing steps from text typesetting, such as merging hyphenated words at line endings or removing marks for highlighting (cf. Grefenstette/Tapanainen 1994). The same is true for mis-segmentation originating from OCR processes.

In summary, DeReKo tokenization requires fast, accurate and consistent processing of heterogeneous data for different linguistic application purposes on commodity hardware. The tools to consider should be easily extensible and maintainable. Both the scenario outlined

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

here and the evaluation of state-of-the-art approaches represent only a snapshot. At the same time it is an update and an extension of similar studies on tools for different NLP tasks in German (Ortmann/Roussel/Dipper 2019), but with a focus on tokenization. In the future, evaluations will be carried out with regard to changing scenarios and new developments in corpus technology.

4. Evaluation

Comparing different software often brings up issues, especially in a task area that is not clearly defined, as tokenization is. As mentioned at the beginning, the scenarios for which tokenization is used differ considerably, which has an influence on the design and applicability of the software. In fact, comparing off-the-shelf solutions with tailor-made tools is always unfair. Therefore, it should be clearly stated that the following comparisons refer to the presented scenario only. In addition, secondary functions of the different tools are not taken into account, like the return of token classes or normalization, even though they can be essential in other scenarios and can have a negative impact especially with regard to speed. The tools are only tested via command line interfaces, so we do not take into account different programming languages. If no native command line tool exists, we have written minimal wrappers following instructions, which should be taken into account regarding speed comparisons as well. We also consider the tools only with respect to a single language (German), while many tools have a primary focus on cross-language applicability. Furthermore, the hardware and software architecture used has a strong influence on the results. As mentioned in Section 3, our tests disfavour applications that can use dedicated GPU support, for example. We provide the full test suite in the form of a Dockerfile[1] to make it replicable for other users and on other systems. Table 2 (at the end of this article) provides an overview of our evaluation for all tools in the categories presented.

4.1 Tools

The evaluated tools are all available under different open licenses and represent in our eyes the state of the art with respect to tokenization and sentence segmentation for German, although we cannot claim to be exhaustive.

Our tools for token and sentence boundary detection

- **KorAP-Tokenizer**[2] is rule-based and compiles, using the lexical analysis generator framework JFlex,[3] a list of regular expressions into a deterministic finite state automaton that can introduce segment boundaries at terminal nodes. The ruleset is based on Apache Lucene's tokenizer and has been extensively modified. Rulesets are available for English, French and German. KorAP-Tokenizer is used productively for tokenization and (among other tools) for sentence segmentation of DeReKo.
- **Datok**[4] (Diewald 2022) is rule-based and generates an extended deterministic finite state automaton based on a reduced finite state transducer generated by XFST (Beesley/Kart-

[1] https://github.com/KorAP/Tokenizer-Evaluation.

[2] https://github.com/KorAP/KorAP-Tokenizer.

[3] https://jflex.de/.

[4] https://github.com/KorAP/Datok.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

tunen 2003). The ruleset of KorAP-Tokenizer was translated to XFST for this purpose. The generation is done with Foma (Hulden 2009). Rulesets are only available for German at this time. Datok is currently being evaluated experimentally.

Tools for token and sentence boundary detection

- **BlingFire**[5] is rule-based and compiles a deterministic finite state automaton based on regular expressions, which segments at terminal nodes. The tested model is implemented cross-language with a focus on English.
- **Cutter** (Graën/Bertamini/Volk 2018) is rule-based and recursively applies language-specific and language-independent rules to a text to segment it. Compared to other rule-based tools, Cutter uses a context-free rather than a regular grammar.
- **JTok**[6] is based on cascading regular expressions that segment tokens until they can be assigned to a token class. Rules exist for English, German and Italian.
- **OpenNLP**[7] is a framework that offers tokenizers and sentence segmenters in different models, both based on maximum entropy. In addition, OpenNLP offers SimpleTokenizer, a tool based on simple character class decisions.
- **SoMaJo** (Proisl/Uhrig 2016) is rule-based and applies a list of regular expressions to segment a text. SoMaJo won first place in the competition of the aforementioned EmpiriST 2015 Shared Task for tokenizing German language Web and CMC corpora and has been regularly improved since then. SoMaJo is available specifically for German.
- **SpaCy**[8] is a framework in which the tokenization stage is rule-based and runs in several phases in which the tokens are split into increasingly finer segments. Rulesets are provided for numerous languages. Different models are offered for sentence segmentation: *Sentencizer* is rule-based, *Dependency* performs a syntactic analysis, *Statistical* segments based on a simple statistical model.
- **Stanford Tokenizer**[9] is rule-based, and relies on JFlex (see KorAP-Tokenizer) to compile a deterministic finite state automaton based on a list of regular expressions that can introduce segment boundaries at terminal nodes.
- **Syntok**[10] is rule-based and applies successive separation rules, primarily in the form of regular expressions, to an input string for segmentation. There is both a tokenizer and a sentence segmenter based on it. Syntok was the fastest tokenizer in Ortmann/Roussel/Dipper (2019). Rules exist for Spanish, English, and German.
- **Waste** (Jurish/Würzner 2013) is based on a hidden Markov model in which a pre-segmented stream of (pseudo)tokens are re-evaluated at the boundaries found and classified as to whether they are word-initial or sentence-initial.

5 https://github.com/microsoft/BlingFire.

6 https://github.com/DFKI-MLT/JTok.

7 https://opennlp.apache.org/.

8 https://spacy.io/.

9 https://nlp.stanford.edu/software/tokenizer.shtml.

10 https://github.com/fnl/syntok.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Tools for token boundary detection only

- **Elephant**[11] (Evang et al. 2013) is an ML system for segmentation based on Conditional Random Fields and Recurrent Neural Networks. We evaluate here a wrapper implementation[12] (Moreau/Vogel 2018) that considers only token segmentation and not sentence segmentation, although Elephant provides both.
- **TreeTagger** (Schmid 1994) is a part-of-speech tagger that carries a separate rule-based tokenization tool that also uses a set of regular expressions to segment a text. The tokenizer does not itself introduce markers for sentence boundaries.

Tools for sentence boundary detection only

- **Deep-EOS** (Schweter/Ahmed 2019) is based on different implementations of neural networks with long short-term memory (LSTM), bidirectional LSTM, and convolutional neural networks. It is not based on pre-tokenization and operates directly on character streams.
- **NNSplit**[13] is an ML approach based on a byte-level LSTM neural network.

In the list of tools we compare here, it is striking that rule-based procedures still dominate tokenization even in modern frameworks, although this is decreasing in other areas of NLP. For sentence boundary recognition, on the other hand, ML techniques seem to be slowly replacing rule-based procedures in this area. ML methods have experienced an increase in importance in recent years due to the availability of large corpora and more powerful computers. But deterministic methods have also benefited (albeit to a lesser extent), through the efficient application of arbitrarily large rulesets and almost arbitrarily large lexicons.

4.2 Performance: speed

Tokenization is not only an NLP problem that can be considered relatively simple, but also one that takes little time to process (compared to, e. g., syntactic parsing). Therefore, when evaluating new tokenizers for research, it is uncommon to specify the runtime. This is slowly changing in the context of machine learning, where tokenizers are used as a pre-processing step for training with very large data sets and speed is therefore of greater importance. But when processing very large corpora in a research context, runtime must be taken into account as well.

For the benchmarking, the novel "Effi Briest" by Theodor Fontane in the Project Gutenberg version was used (with a total of 98,207 tokens).[14] The measures correspond to the average value of 100 runs. Since the length of a text can have an impact on performance, a tenfold concatenation of the text was also tested.

Figure 1 compares the speed of all tools we measured in terms of "tokens per millisecond".[15] Detailed values are listed in Table 1.

[11] https://gmb.let.rug.nl/elephant/.

[12] https://github.com/erwanm/elephant-wrapper.

[13] https://bminixhofer.github.io/nnsplit/.

[14] Based on the information provided by the Unix tool 'wc -w'.

[15] The test system is an Intel Xeon CPU E5-2630 v2 @ 2.60GHz with 12 cores and 64 GB of RAM.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

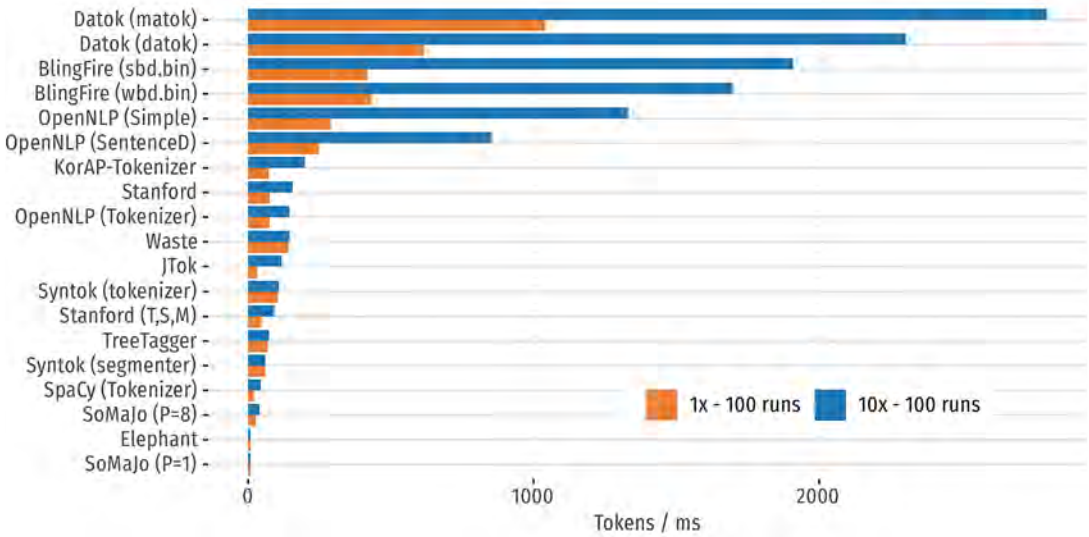


Fig. 1: Speed comparison of all tested tools with different models and configurations

With respect to DeReKo, these times can be extrapolated for individual cores. Our experimental tool Datok (Model "matok") is the fastest tool in the comparison and could completely tokenize and sentence segment DeReKo within ~13.5h,[16] closely followed by Bling-Fire with ~33h. KorAP-Tokenizer currently takes ~193h for the same task, just over 8 days. Most other rule-based off-the-shelf tools for tokenization are also in the same range. SoMaJo, which is to be emphasized with respect to German CMC data (s. sec. 4.3), would require about 72 days (without direct parallelization).[17] Cutter was not able to fully segment the large batch size and would have taken just over 4 years to segment DeReKo at small batch sizes. In particular, Deep-EOS and NNsplit show that their use without dedicated hardware (GPU) is not an option for sentence segmentation in our scenario: full processing of DeReKo would take between 1.7 and almost 6.4 CPU years on the hardware used, with this being in addition to tokenization. NNSplit admits to poor performance in terms of CPU-only usage and claims to be twice as fast as SpaCy's Sentencizer in GPU usage[18]. A similar speed increase should be expected with Deep-EOS.

It can be seen that the speed of tokenization and sentence segmentation is a variable to be considered for large data sets, and not all tools are capable of complete data resegmentation in an acceptable amount of time. It should be noted though that even slower tokenizers like Elephant and SoMaJo (on one core) can still process over 8,000 tokens in one second – and are perfectly adequate for most scenarios and corpus sizes.

4.3 Performance: quality

Even though processing speed is the focus of our review, a high quality of tokenization is of primary importance in the aforementioned scenario. The evaluation of tool quality in an NLP context is typically operationalized via a comparison to a Gold Standard. Already in the seemingly simple case of tokenization, however, the definition of what a token is and where

[16] All extrapolations are based on the measured values of "1 x Effi" in Table 1.

[17] SoMaJo supports the use of multiple processor cores, which we have included in our benchmarks, but is less of an issue, as parallelization happens at a higher level in our scenario.

[18] https://bminixhofer.github.io/nnsplit/#benchmark

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

its boundaries are leaves room for interpretation (s. sec. 2). Therefore, it is not appropriate to speak of "correct" or "incorrect" for a tokenization – it often depends on the field of application what makes a token an independent entity (s. sec. 3). This has to be taken into account for evaluation related to existing corpora that may follow different guidelines. The same holds for sentence boundaries.

We use the EmpiriST Web and CMC corpora as well as version 2.9 of the German Universal Dependency GSD Corpus (McDonald et al. 2013) in our evaluation of tokenization. We also use the latter for evaluating sentence boundary detection. We rely on the EmpiriST software[19] as the tool for computing F_1 values.

Token boundary detection

All tools except the OpenNLP Simple Tokenizer achieve an F_1 score well above 99% for the UD corpus (s. Tab. 1). For the web corpus, SoMaJo, Cutter, TreeTagger, KorAP-Tokenizer and Datok also achieve more than 99%. With respect to the CMC corpus, SoMaJo, Stanford Tokenizer, TreeTagger, Cutter, KorAP-Tokenizer, and Datok achieve an F_1 score above 95%. Regarding our preference to use only one tool for different corpora, we consider them more suitable for our purposes. Nonetheless, the evaluation of quality is not very meaningful with respect to KorAP-Tokenizer and Datok, since rule-based approaches can be optimized for the evaluation data, and thus in many cases perfect accuracy can be achieved (with limitations, s. sec. 4.6). Accordingly, they cannot be compared with the evaluation within EmpiriST, which was about testing against unknown data.

Tool	V.	Model	UD-GSD (Tokens)	Empirist-CMC	Empirist-Web	UD-GSD (Sententces)	1 x Effi	10 x Effi
			F_1	F_1	F_1	F_1	T/ms	T/ms
KorAP-Tokenizer	2.2.2		99.45	99.06	99.27	96.87	72.90	199.28
Datok	0.1.5	datok	99.45	98.79	99.21	97.60	614.72	2304.13
		matok					1041.63	2798.78
BlingFire	0.1.8	wbd.bin	99.25	55.85	95.80	-	431.92	1697.73
		sbd.bin	-	-	-	95.90	417.10	1908.87
Cutter	2.5		99.47	96.24	99.38	97.31	0.38	-
JTok	2.1.19		99.56	58.44	98.09	97.92	31.19	117.22
OpenNLP	1.9.4	Simple	95.70	55.26	91.69	-	290.71	1330.23
		Tokenizer (de-ud-gsd)	99.67	65.22	97.58	-	74.65	145.08
		SentenceDetector (de-ud-gsd)	-	-	-	98.51	247.84	853.01
SoMaJo	2.2.0	p=1	99.46	99.21	99.87	97.05	8.15	8.41
		p=8					27.32	39.91

[19] The comparison tool was developed by Stephanie Evert and published under GPL v3. For the evaluation of the sentence boundaries, the segmented sentences are taken instead of single tokens.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Tool	V.	Model	UD-GSD (Tokens)	Empi-rist-CMC	Empi-rist-Web	UD-GSD (Sent-entces)	1 x Effi	10 x Effi
			F_1	F_1	F_1	F_1	T/ms	T/ms
SpaCy	3.2.3	Tokenizer	99.49	69.94	98.29	-	19.73	44.40
		Sentencizer	-	-	-	96.80	16.94	40.58
		Statistical	-	-	-	97.16	4.90	10.01
		Dependency	-	-	-	96.93	2.24	0.48
Stanford	4.4.0	tokenize	99.93	97.71	98.46	-	75.47	156.24
		tokenize,ssplit, mwt				98.22	46.95	91.56
Syntok	1.4.3	Tokenizer	99.41	70.76	97.50	-	103.90	108.40
		Segmenter	-	-	-	97.50	59.66	61.07
Waste	2.0.20-1		99.55	65.90	98.49	97.46	141.07	144.95
Elephant	0.2.3		99.62	66.96	97.88	-	8.57	8.68
TreeTag-ger	3.2.4		99.52	95.58	99.27	-	69.92	72.98
Deep-EOS	0.1	bi-lstm-de	-	-	-	97.47	0.25	0.24
		cnn-de	-	-	-	97.49	0.27	0.25
		lstm-de	-	-	-	97.47	0.29	0.27
NNSplit	0.5.8		-	-	-	95.55	0.90	0.90

Table 1: Overview of all compared tools and models with their performance measures (best three highlighted in each category)

Moreover, the quality with respect to the applied machine learning tools says less about the implementation or the algorithm than about the corpus used for training. Thus, we compare here "off-the-shelf" solutions with default configurations for the outlined scenario without checking whether a tool trained and adjusted for our purposes would achieve better results.

Taking this expected bias into account, the results show that high tokenization speed does not require any compromises in terms of quality. Furthermore, the comparison of the three evaluated datasets shows that approaches exist that work across genres and thus facilitate the use in the outlined scenario.

Sentence boundary detection

All tools show F_1 values above 95% and can therefore be considered suitable for sentence boundary detection. Stanford and the OpenNLP model show values above 98%. They were also developed and trained using the test corpus. Our tools KorAP-Tokenizer and Datok perform weaker in relation – a readjustment is desirable.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

4.4 Extensibility and adaptability

Machine learning-based approaches have the advantage of being easily transferable to other languages or genres in the presence of appropriately annotated corpora. "This is why [Moreau and Vogel] argue that the evaluation of software tools should progressively shift the focus from accuracy in a specific language to robustness and adaptability to a wide range of languages." (Moreau/Vogel 2018, p. 1119). Also, with respect to novel linguistic phenomena (as they occur in various CMC corpora mentioned above), rule-based approaches are in principle inferior, since they cannot deal with unknown phenomena to begin with. However, SoMaJo's win at EmpiriST 2015 shows that a rule-based system designed for extensibility nevertheless offers advantages that can make it competitive in many scenarios. Graën/Bertamini/Volk (2018) even put their main focus regarding their rule-based system Cutter on extensibility and on iterative adaptation to new languages.

The good adaptability of rule-based systems with respect to new language phenomena is largely due to the pattern-like nature of these entities, such as email addresses, emoticons, or XML fragments. These can be well formulated in rule-based terms. Extralinguistic units also occur in other word-segmented languages (Graën/Bertamini/Volk 2018), which makes them *language-independent*. *Language-specific* rules, on the other hand, include general definitions of words and how to deal with punctuation, especially in sentence segmentation. In addition, there are usually lists of common abbreviations (for the correct treatment of periods, e. g. "etc.") and known proper names with non-alphabetic symbols (e. g. "3G+") in that language. As mentioned in Section 4.1, both KorAP-Tokenizer and Datok are based on widely used and well documented frameworks for the rule definition of lexical analysers. Both systems can be adapted with little training, especially with regard to the various lists that can be extended without any knowledge of syntax. By reusing the language-independent rules, the approaches can also be adapted for other languages with manageable effort. This has already been done for KorAP-Tokenizer with respect to English and French.

4.5 Reproducibility and maintainability

A distinct advantage of rule-based systems over machine-learning approaches, especially in the scientific context, is the reproducibility of tokenization. Thus, errors can be traced back to individual rules, which can be easily and likewise systematically corrected, validated with reference to regression tests (cf. Graën/Bertamini/Volk 2018), and versioned. Corrections can be made consistently across the data. Trained machine learning approaches have the disadvantage that an extended corpus must first be created in order to (re-)train the tools. And only very extensive and difficult-to-maintain tests can ensure that re-training does not cause regressions. Rule-based approaches therefore allow for a good balance between correctness and reproducibility in the processing of scientific research data.

In terms of maintainability, data consistency, and runtime, it is also advantageous if token and sentence segmentation can be performed with a single tool, based on the same model.

Also of importance for maintainability are short development cycles. Training and testing machine-learning methods is time-consuming and can often be solved in acceptable time only with special hardware. Compiling complex finite state automata of rule-based approaches can also lead to significant time and resource consumption. Systems like SoMaJo or Cutter can be tested without intermediate steps while providing good extensibility, which simplifies their maintenance. Separating the language model from program code also facil-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

itates maintenance and versioning. In the case of machine-learning, this is usually the case, but rule-based approaches also often have separate language models, for example Cutter, but also (in part) the Jflex-based tools such as Stanford-Tokenizer and KorAP-Tokenizer, and Datok's XFST model.

Tool	Tokens & Sentences	Speed	Quality	Extensibi-lity	Adaptabi-lity	Maintai-nability	Reprodu-cibility
KorAP-Tokenizer	•••	••	•••	•••	••	••	•••
Datok	•••	•••	•••	•••	•	••	•••
BlingFire	••	•••	•	••	••	••	•••
Cutter	•••	-	•••	••••	••••	••••	•••
JTok	•••	••	•	••	•	••	•••
OpenNLP	••	••	•	•	•••	•	-
SoMaJo	•••	•	••••	•••	•	••••	•••
SpaCy	••	•	•	•••	•••	••	••
Stanford	••	••	•••	•••	••	••	•••
Syntok	•••	••	•	•••	••	•	•••
Waste	•••	••	•	•	••	•	-
Elephant	•	•	•	•	••	•	-
TreeTagger	•	••	•••	•••	•	•••	•••
Deep-EOS	•	-	••	•	•••	••	-
NNSplit	•	-	•	•	•••	••	-

Table 2: Overview of all compared tools with ratings in the examined categories

4.6 Limitations

There are some fundamental limitations regarding single-pass finite state systems that need to be considered, nonetheless, and that apply to our approaches.

Graën/Bertamini/Volk (2018) primarily point out long-distance relationships between tokens as a weakness of finite state based systems and as an example refer to the use of the apostrophe in German as a possessive marker for words ending in a phonetic /s/. These apostrophes must belong to the preceding token in the possessive case, but can also represent the end of an expression in simple quotation marks. Without the context of a starting quotation expression, a decision in a finite state-based approach is not possible – accordingly, neither KorAP-Tokenizer nor Datok can make these decisions.

Furthermore, the strict left-longest-match directive means that sometimes valid tokens can never be segmented. An example would be (following general word and URL rules) the string "Go tohttp://google.com/", in which a space was omitted by mistake and which would currently not be segmented into the expected tokens "Go", "to" and "http://google.com/" by both KorAP-Tokenizer and Datok, since "tohttp" is considered a valid word token and is a longest-match accordingly. The subsequent tokens would be further segmented as the URL rule would not apply. More limitations concern the processing of emoji sequences, which are difficult to represent in strict finite-state models without character ranges.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

5. Summary

Tokenization and sentence boundary detection for texts of word-segmented writing systems belong to the simpler and faster tasks of NLP, and already with naïve approaches good results can be achieved. However, tokenization errors can have a cascading effect on further analysis steps, which is why high quality is of great importance. The emergence of new token types and very large data volumes in the context of unedited, heterogeneous CMC corpora pose further new challenges, especially in research data processing.

In this paper, different tokenizers are compared with respect to this scenario for German in terms of their processing speed, quality, extensibility, adaptability, maintainability and reproducibility. We also present the approaches that are being pursued in building DeReKo.

We believe that processing speed for very large data is a dimension that should not be neglected and that approaches are possible that do not have to compromise on quality with respect to heterogeneous data. In our opinion, the maintainability and reproducibility of process results of rule-based systems represent a further advantage over machine-learning approaches (at least currently), especially in the context of research data processing. But this assessment is only a snapshot: Future developments in this area as well as changes to the scenario described make constant re-evaluations necessary.

We would like to acknowledge the helpful comments of the anonymous reviewers, which clearly contributed to the improvement of the present study.

References

Bartz, T./Beißwenger, M./Storrer, A. (2013): Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation: Phänomene, Herausforderungen, Erweiterungsvorschläge. In: JLCL 28 (1).

Beesley, K. R./Karttunen, L. (2003): Finite state morphology. (= CSLI Studies in Computational Linguistics). Stanford.

Beißwenger, M./Bartsch, S./Evert, S./Würzner, K.-M. (2015): Richtlinie für die manuelle Tokenisierung von Sprachdaten aus Genres internetbasierter Kommunikation. Guideline document from the Empirikom shared task on automatic linguistic annotation of internet-based communication. In: EmpiriST 2015.

Beißwenger, M./Bartsch, S./Evert, S./Würzner, K.-M. (2016): EmpiriST 2015: A shared task on the automatic linguistic annotation of computer-mediated communication and web corpora. Proceedings of the 10th Web as Corpus Workshop, pp. 44–56.

Chiarcos, C./Ritz, J./Stede, M. (2009): By all these lovely tokens... merging conflicting tokenizations. Proceedings of the Third Linguistic Annotation Workshop (LAW III), pp. 35–43.

Diewald, N. (2022): Matrix and double-array representations for efficient finite state tokenization. In: Proceedings of the 10th Workshop on Challenges in the Management of Large Corpora (CMLC-10) at LREC 2022. Marseille.

Evang, K./Basile, V./Chrupała, G./Bos, J. (2013): Elephant: sequence labeling for word and sentence segmentation. Proceedings of the EMNLP 2013: Conference on Empirical Methods in Natural Language Processing. Seattle.

Graën, J./Bertamini, M./Volk, M. (2018): Cutter – a universal multilingual tokenizer. In: Cieliebak, M./Tuggener, D./Benites, F. (eds.): Swiss text analytics conference, Nr. 2226, pp. 75–81.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Grefenstette, G./Tapanainen, P. (1994): What is a word, What is a sentence? Problems of tokenization. Proceedings of COMPLEX '94, pp. 79–87.

Hulden, M. (2009): Foma: a finite-state toolkit and library. Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics, pp. 29–32.

Jurish, B./Würzner, K.-M. (2013): Word and sentence tokenization with Hidden Markov Models. JLCL, 28 (2), pp. 61–83.

Kupietz, M./Lüngen, H./Kamocki, P./Witt, A. (2018): The German Reference Corpus DeReKo: New developments – new opportunities. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), pp. 4353–4360.

McDonald, R./Nivre, J./Quirmbach-Brundage, Y./Goldberg, Y./Das, D./Ganchev, K./Hall, K./Petrov, S./Zhang, H./Täckström, O./Bedini, C./Bertomeu Castelló, N./Lee, J. (2013): Universal dependency annotation for multilingual parsing. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, pp. 92–97.

Moreau, E./Vogel, C. (2018): Multilingual word segmentation: training many language-specific tokenizers smoothly thanks to the Universal Dependencies Corpus. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki.

Ortmann, K./Roussel, A./Dipper, S. (2019): Evaluating off-the-shelf NLP tools for German. In: Proceedings of the 15th conference on natural language processing (KONVENS 2019). Erlangen, pp. 212–222.

Palmer, D. D./Hearst, M. A. (1997): Adaptive multilingual sentence boundary disambiguation. In: Computational Linguistics, 23 (2), pp. 241–267.

Proisl, T./Uhrig, P. (2016): SoMaJo: state-of-the-art tokenization for German web and social media texts. Proceedings of the 10th Web as Corpus Workshop, pp. 57–62.

Schmid, H. (1994): Probabilistic part-of-speech tagging using decision trees. Proceedings of International Conference on New Methods in Language Processing.

Schweter, S./Ahmed, S. (2019): Deep-EOS: general-purpose neural networks for sentence boundary detection. Proceedings of the 15th Conference on Natural Language Processing (KONVENS). Erlangen.

Webster, J. J./Kit, C. (1992): Tokenization as the initial phase in NLP. Proceedings of the 14th Conference on Computational Linguistics (4), pp. 1106–1110.

Contact information

Nils Diewald
Leibniz-Institut für Deutsche Sprache
diewald@ids-mannheim.de

Marc Kupietz
Leibniz-Institut für Deutsche Sprache
kupietz@ids-mannheim.de

Harald Lüngen
Leibniz-Institut für Deutsche Sprache
luengen@ids-mannheim.de

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Ana-Maria Gînsac/Mihai-Alex Moruz/Mădălina Ungureanu

THE FIRST ROMANIAN DICTIONARIES (17TH CENTURY). DIGITAL ALIGNED CORPUS

Abstract This paper presents the project "The first Romanian bilingual dictionaries (17th century). Digitally annotated and aligned corpus" (eRomLex) which deals with the editing of the first bilingual Romanian dictionaries. The aim of the project is to compile an electronic corpus comprising six Slavonic-Romanian lexicons dating from the 17th century, based on their relatedness and the fact that they follow a common model in order to highlight the characteristics of this lexicographical network (the affiliations between the lexicons, the way they relate to the source, the innovations towards it, their potential uses) and to facilitate the access to their content. A digital edition allows exhaustive data extraction and comparison and link with other digitized resources for old Romanian or Church Slavonic, including dictionaries. After presenting the corpus, we point to the necessary stages in achieving this project, the techniques used to access the material and the challenges and obstacles we encountered along the way. We describe how the corpus was created, stored, indexed and can be searched over; we will also present and discuss some statistical analyses highlighting relations between the Romanian lexicons and their Slavonic-Ruthenian source.

Keywords Romanian lexicography; 17th century; Church Slavonic; bilingual dictionaries in electronic format

1. Introduction

Our study focuses on the first Romanian bilingual dictionaries (six Slavonic-Romanian dictionaries dating from the 17th century) and their digital editing. Starting from the context of their elaboration, we present the stages of the digitization project, also pointing at how this edition can be exploited and integrated into other language resource projects dedicated to Romanian.

1.1 Context

The first half of the 17th century in Eastern Europe saw the growing expansion of Catholicism and Protestant currents in an area dominated by Orthodoxy. This expansion triggered various reactions from the Orthodox clergy. One of the most notable such reactions came from the Metropolitan of Kiev, Petru Movilă (1597–1647), who translated and elaborated works related to worship and dogma. The issue of a series of Church Slavonic linguistic instruments, among which Pamvo Berynda's Slavonic-Ruthenian Lexicon (1627) and Meletie Smotrițki's Grammar (1619), was also associated with the context of this counter-reformation of Orthodoxy.

Petru Movilă's descent from a noble Romanian family played a key part in the close cultural relations between Kiev and the Romanian Principalities (Moldova and Wallachia), and thus his activity had a considerable influence on the cultural movement in the Romanian Principalities. The editing and copying of Slavonic-Romanian bilingual lexicons, having as a model Berynda's lexicon, could be associated to the same context. They are the most consistent part of the Romanian lexicography before the 18th century (for an overview, see Seche 1966, pp. 9–11).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

1.2 Corpus

The six Slavonic-Romanian lexicons compiled in the second half of the 17th century having as a model Pamvo Berynda's Lexicon (= Lex.Ber.) are edited within the eRomLex project (see infra, 2); they are large (Lex.Ber. has more than 7,000 entries and the inventories of the Romanian dictionaries are equally extended, except for Lex.Mard., which records 4,574 entries), are complete (a letter is missing from one of them and some pages from another), all preserved in manuscripts. Four of these are kept at the Romanian Academy Library in Bucharest and two can be found in Russian Libraries (in Moscow and Sankt Petersburg). They all date from the same era, come from the same area (today's Wallachia) and seem (probably with one exception) related to each other not only by the common Slavonic model, but as modified copies of a unique Romanian model, either lost or still undiscovered (see Felea 2021). Lex.St. and Lex.3473 are part of miscellanies that comprise a Romanian version of Meletie Smotrițki's grammar; this fact also gives an idea of the purpose of their compiling. These lexicons are: the Lexicon from Rom. ms. no. 1348, Library of the Romanian Academy, Bucharest (1-84v) (furthermore Lex.1348); the Lexicon from Rom. ms. no. 3473, Library of the Romanian Academy, Bucharest (files 1-369) (= Lex.3473); the Lexicon of Mardarie Cozianul, Rom. ms. no. 450, Library of the Romanian Academy, Bucharest, 1649 (= Lex.Mard.); the Lexicon from Moscow, Russian State Archive of Old Documents, Fond 188, Оп. 1. Ч. 2., p. 491 (= Lex.Mosc.); the Lexicon from Petersburg, the National Library of Russia in Sankt Petersburg (notice n° Q.XVI.5 – Славяно-молдавскийсловарь) (= Lex.Pet.); The Lexicon from Rom. ms. no. 312, Library of the Romanian Academy, Bucharest (41r-216v) (= Lex.St.). For a more thorough description of the lexicons, see Gînsac/Ungureanu (2018, pp. 850–853). The term corpus, as used throughout this paper, is not used in the more common sense given in corpus linguistics, but in the broader sense of collections of writings having something in common. In the specific context of dictionaries, most of the text included does not contain proper sentences, but rather disparate words or word lists, given as glosses for Slavonic headwords. Because of this, we consider that the absolute number of words is less relevant than the number of entries (see section 3).

1.3. Current state of knowledge

1.3.1 Concerning the research on the lexicons

This group of lexicons has been studied rather sporadically; the only edited lexicon is the oldest one, Lex.Mard. (Crețu 1900); the others have been analysed focusing on small samples. Crețu (1900), in his introductory study, provides a brief description of the whole group; Bogdan (1891) focuses on the description of Lex.Pet., for which he also establishes the source; Ciobanu (1914) deals in the same manner with Lex.Mosc. This issue has been revisited recently (Gînsac/Ungureanu 2018; Felea 2021), and the comparative editing of the six lexicons was proposed as an objective of the eRomLex project: "The first bilingual Romanian dictionaries (the 17th century). Digitally annotated and aligned corpus".[1]

1.3.2. Electronic lexicographic corpora

The earliest efforts towards the digitization of Romanian lexicographic resources targeted modern dictionaries. The Romanian Language Thesaurus (Romanian Academy, 1906–2010,

[1] See Gînsac/Moruz/Ungureanu 2021; http://www.scriptadacoromanica.ro/bin/view/eRomLex/.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

19 volumes) is available within the eDTLR project (Romanian Language Dictionary in Electronic Format) and provides multiple interrogation criteria based on: word, form, etymology, frequency, age, dialectal area, stable combinations, compound words, authors, period (Cristea et al. 2011). The main objective of the more recent CLRE project (Essential Romanian Lexicographic Corpus) is the alignment at the entry-level of the most important dictionaries – old and new, general or specific – of the Romanian language (for further details, see Clim 2015, pp. 101–104).

The Multilingual Buda Lexicon (1825), considered the first modern normative dictionary of the Romanian language, was edited and processed in electronic format between 2011 and 2013 (available at: https://doi.org/10.26424/lexiconuldelabuda). The dictionary can be consulted according to several criteria: lemma, language (Romanian, Latin, Hungarian and German), grammatical form, semantic-stylistic value, etymology, idioms, phrases, quotations etc.

In South-Eastern Europe there are some old multilingual lexicographical resources in digital format, of great importance for the Romanian language, which used the Cyrillic alphabet until the 19th century. F. Miklosich's Lexicon Palaeoslovenico-Graeco-Latinum is regarded as the most relevant dictionary in Slavic studies, and its digital edition allows interrogation by words, word parts and grammatical categories (Miklosich 1865).

The Old Church Slavonic Dictionary is available in digital format for Bulgarian (Totomanova 2021), allowing word-based interrogation (see https://histdict.uni-sofia.bg/oldbgdict/oldbg_search/). This dictionary is part of a digital platform for Bulgarian language and literature containing: Unicode fonts, diachronic corpora, historical dictionaries equipped with tools for writing and editing the entries, grammatical dictionaries, prototypical search engine, and virtual keyboard (Ganeva 2018, p. 117; see also Tasovac 2020).

A similar digital resource for Slavonic is Gorazd: An Old Church Digital Hub (http://www.gorazd.org/?q=en/node/12), a database developed by the Institute of Slavic Studies (The Czech Republic Science Academy), which includes three modern dictionaries of Church Slavonic (Knoll 2021). The database is accessible via the "Gulliver" interface translated into Czech and English. The interface can be interrogated by: dictionary, word, grammatical ending, type of entry (main, cross-reference, exhausted), texts quoted within the entries.

2. Project presentation

2.1. Corpus creation and storage

The first stage in creating of the corpus consisted of obtaining the lexicons in an editable format. To this end, we tested Transkribus, an automatic handwriting recognition programme (see https://readcoop.eu/transkribus/). However, the specific format of the lexicons (written on columns), particular handwriting (in some cases quite irregular) made the use of Transkribus relatively inefficient (see Fig. 1; manual correction would have been extremely time consuming), therefore the editable format was obtained manually.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

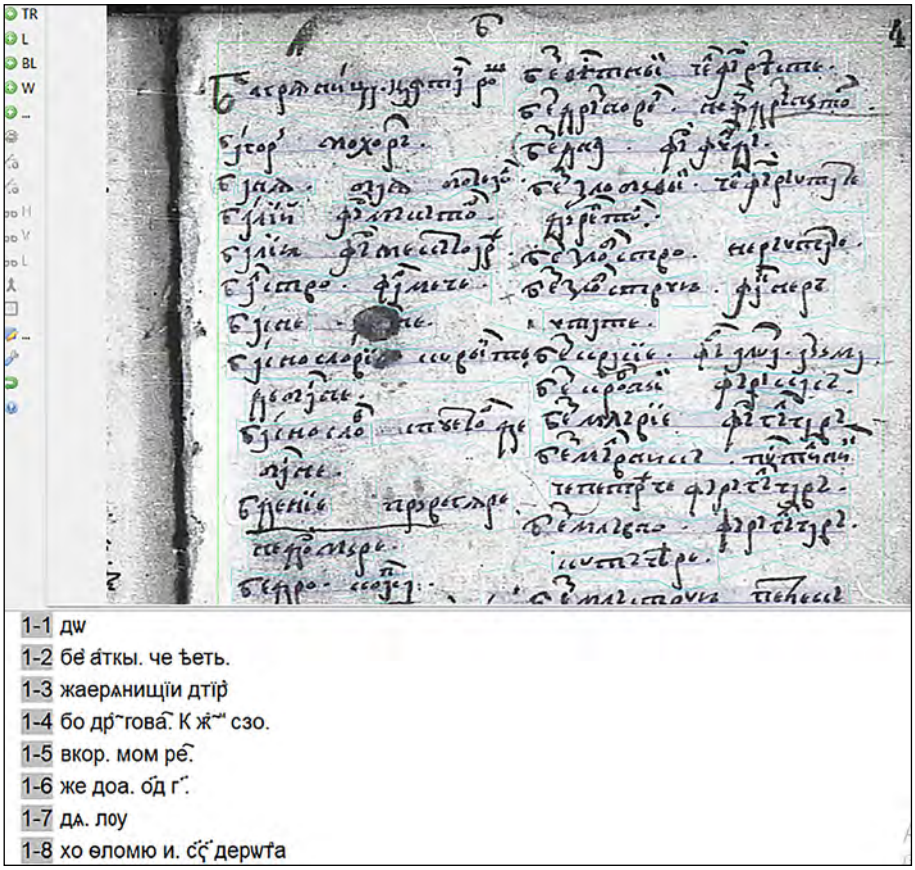


Fig. 1: Automatic text recognition using Traskribus on part of Lex.St.

Each entry was filed in a standard Word form, allowing users to automatically extract distinct information (headword, definition, and location: row, column, entry number for recording the sequence of entries, comments).

2.2 Indexing

From the standard form discussed above, we have automatically extracted the entries in a structured format to be imported into the platform. To this end, we have transformed the Word form to an XML format (MS Word XML) through Visual Basic scripts. This XML format, however, is complex and large amounts of formatting information are not relevant to the extraction process. We have thus transformed the Word XML format to a more simplified version, which contains only the necessary information. In order to better control the flow of imported entries (so as to spot potential errors in the processing chain, for example), we chose to perform the processing at the letter and volume level: we only import the entries beginning with the same letter from all the lexicons.

After the conversion to a structured format, the entries were automatically aligned. This was done manually by pointing to the equivalent Lex.Ber. entry (if extant) during the transcription process. To reduce alignment error rate, we have built, on the basis of the Lex.Ber. equivalent, a simplified form of the word by removing of accents, replacing of superscripts with standard letters, and replacing equivalent letters with the same surface representation (e. g., the letters ї ї ı и й І И are transformed into и). Replacement of letters is done according to a correspondence matrix, developed iteratively, which, at the time of writing, contains

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

more than 200 replacement pairs. This simplified form presents multiple advantages: alignment is more precise, as spelling differences are flattened; words which are not found in Lex.Ber. can be aligned as well; searching is easier and more intuitive over simplified forms, since user cannot know what writing peculiarities a given word might have. The aligned entries are then stored as one object, with multiple equivalent entries from different lexicons, in a proprietary XML format. This format can be easily transformed to XML TEI or PDF by means of XSL transformations.

For storing the entries in an indexed form, and for efficient and intuitive searching and viewing, we have chosen the XWiki[2] technology. This allows for collaborative editing of entry sets, fast and parameterizable searching (via scripting and plugins), layered access and restricting permissions so as to avoid accidental modifications.

2.3 Searching the corpus

In XWiki, each entry is a Wiki page, which is automatically indexed in the platform database, and can be efficiently interrogated using much faster tools than XQuery, XWiki Query Language, which is significantly more efficient as it is derived from SQL and runs on a specific instance of a relational database. For better accuracy of search, we have also lemmatized and performed morphological analysis on the glosses, using the tool described in (Patraș/Pavel/Haja 2007).

Title	Location
възра	
възрасть	Lexicon / възрасть
възражаю	Lexicon / възражаю
възрастаю	Lexicon / възрастаю
възрастеть	Lexicon / възрастеть
възрастет	Lexicon / възрастет
възрастетъ	Lexicon / възрастетъ
възразѝ	Lexicon / възразѝ
възрастъ	Lexicon / възрастъ
възраченный	Lexicon / възраченный
възрастѝ	Lexicon / възрастѝ
възраженъ	Lexicon / възраженъ
възражение	Lexicon / възражение

Results 1 - 12 out of 12

Fig. 2: Search interface

[2] www.xwiki.org.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

To prototype the query module and to test its efficiency, we have provided, for the project members, a multi-criterial search interface prototype; this allows for word or part of word searches, and, for testing purposes, page author and date of modification. This prototype will be expanded into the full search interface for the corpus. Figure 2 shows the search interface prototype.

3. Selected results

Since the creation of the lexicon corpus is still a work in progress, we have performed statistical analysis on only part of the corpus. Below, we have given the statistical analysis for the entries under the letter L in all 6 of the lexicons, compared to their source:

- 427 aligned entries; of these, only 173 are found in Lex.Ber.
- Lex.Ber. has 216 entries, which means that 43 of them can't be found in any of the six lexicons
- 248 entries in Lex.Mosc
- 217 in Lex.Staicu, of which 13 are doubles (2 separate entries for the same headword) and 1 is triple
- 255 in Lex 3473, of which 7 are doubles
- 213 in Lex.Pet, of which 11 are doubles
- 228 in Lex1348, of which 12 are doubles and 1 is triple
- 136 in Lex.Mard, of which 2 doubles.

Further research will focus on the words from Lex.Ber. missing from the Romanian lexicons, on similarities and differencies between the lexicons in terms of entries inventory and definition structures.

4. Future work

In terms of future research, we intend to investigate the manner in which the lexical inventory of the lexicons is found in other documents of the period which are available in electronic format. Also, on the basis of the alignment with Lex.Ber., we can align the lexicons with similar resources of the period in other languages and with Romanian lexicographic electronic corpora (see above, 1.3.2.). The valorization of the eRomLex corpus will enrich the already existing lexicographic corpora with new meanings, variations in form or morphology, attestations. Upon completion of the project, the electronic dictionaries will be made freely available, in an open-source format, at www.scriptadacoromanica.ro. The search functions will allow researchers to identify new morphological forms and earlier attestations of words which are not yet present in the Romanian Thesaurus. For translation studies, the eRomLex electronic dictionary could be used in those situations where glosses are encyclopaedic, providing samples for the Romanian language and not just single equivalents, and as a bilingual tool for Slavonic translations into Romanian. The dictionary could also be used for studies regarding the history of mentalities or cultural history.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

References

Lex.Ber. = Pamvo Berynda: Леѯиконъ славеноросскїй и именъ тлъкованїе, Kiev, 1627. Edition by V. Nimciuk, 1961. Online: http://litopys.org.ua/berlex/be.htm (last access: 19-05-2022)

Bogdan, I. (1891): Un lexicon slavo-român din secolul XVI. In: Convorbiri literare 25 (3), pp. 193–204

Ciobanu, Ș. (1914): Славяно-румынскій словарь библіотеки Московскаго Общества Исторіи и Древностей No 240. In: Русскій филологическій вестникъ 71 (1), pp. 75–88.

Clim, M. (2015): La lexicografía rumana informatizada: tendencias, obstáculos y logros. In: Domínguez Vázquez, M. J./Gómez Guinovart, X./Valcárcel Riveiro, C. (eds.): Lexicografia de las lenguas románicas. Aproximaciones a la lexicografía moderna y contrastiva, vol. II. Berlin/München/Boston, pp. 95–110.

Crețu, G. (1900): Mardarie Cozianul. Lexicon slavo-românesc și tîlcuirea numelor din 1649. Bucharest.

Cristea, D./Haja, G./Moruz, A./Răschip, M./Pătrașcu, M. (2011): Partial statistics at the end of the eDTLR project–the Thesaurus Dictionary of the Romanian Language in electronic form. In: Zafiu, R./Uşurelu, C./Oprea, H. Bogdan (eds.), Romanian language. Hypostasis of linguistic variation. Acts of the 10th Colloquium of the Chair of Romanian Language, Bucharest, 3–4 December. Bucharest, pp. 213–224.

Felea, I. M. (2021): The Staicu lexicon in relation to lexicons belonging to the Berynda family: orthography and structure. In: Diacronia 13, A181, pp. 1–13.

Ganeva, G. (2018): Electronic diachronic corpus and dictionaries of Old Bulgarian. In: Studia Ceranea 8, pp. 111–119.

Gînsac, A.-M./Ungureanu, M. (2018): La lexicographie slavonne-roumaine au XVIIe siècle. In: Zeitschrift für romanische Philologie 134 (3), pp. 845–876.

Gînsac, A. M./Moruz, M. A./Ungureanu, M. (2021): Slavonic–Romanian lexicons of the 17th century and their comparative digital edition (the eRomLex project). In: Diacronia 14, A192, pp. 1–11.

Knoll, V. (2021): Gorazd: An Old Church Slavonic digital hub and the Romanian Slavonic studies. In: Diacronia 14, A197, pp. 1–9.

Miklosich, Fr. (1865): Lexicon Palaeoslovenico-Graeco-Latinum. Lipsiae. http://www.monumentaserbica.branatomic.com/mikl2/ (last access: 22-03-2022).

Patraş, V. S./Pavel, G./Haja, G. (2007): Resurse lingvistice în format electronic. Biblia 1688. Regi I, Regi II – probleme, soluții. In: Pistol, I. D./Cristea, D./Tufiş, D. (eds.): Lucrările atelierului Resurse lingvistice şi instrumente pentru prelucrarea limbii române, Iași, 14–15 decembrie 2007. Iași, pp. 51–60.

Seche, M. (1966): Schiţă de istorie a lexicografiei române, vol. I: De la origini pînă la 1880. Bucharest.

Tasovac, T. (2020): The historical dictionary as an exploratory tool: a digital edition of Vuk Stefanovic Karadzic's Lexicon Serbico-Germanico-Latinum. Ph. D. thesis. Dublin. http://www.tara.tcd.ie/bitstream/handle/2262/92750/Tasovac.pdf?sequence=1&isAllowed=y (last access: 22-03-2022)

Totomanova, A. M. (2021): Electronic research infrastructure for Bulgarian medieval written heritage: history and perspectives. In: Diacronia 14, A193, pp. 1–9. https://doi.org/10.17684/i14A193en.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Contact information

Ana-Maria Gînsac
Institute of Interdisciplinary Research, Department of Social Sciences and Humanities, "Alexandru Ioan Cuza" University, Iași
anamaria_gansac@yahoo.com

Mihai-Alex Moruz
Faculty of Computer Science, "Alexandru Ioan Cuza" University, Iași
mmoruz@info.uaic.ro

Mădălina Ungureanu
Institute of Interdisciplinary Research, Department of Social Sciences and Humanities, "Alexandru Ioan Cuza" University, Iași
madandronic@gmail.com

Acknowledgements

This work was supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI – UEFISCDI, project number PN-III-P1-1.1-TE -2019-0517, within PNCDI III.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Iztok Kosem

TRENDI – A MONITOR CORPUS OF SLOVENE

Abstract In this paper we present Trendi, a monitor corpus of written Slovene, which has been compiled recently as part of the SLED (Monitor corpus and related resources) project. The methodology and the contents of the corpus are presented, as well as the findings of the survey that aimed to identify the needs of potential users related to topical language use. The Trendi corpus currently contains news articles and other web content from 110 different sources, with the texts being collected and linguistically annotated on a daily basis. The corpus complements Gigafida 2.0, a 1.13-billion-word reference corpus of standard written Slovene. Also discussed are the ways in which the corpus will be integrated into various lexicographic projects, helping not only in the identification of neologisms but also in monitoring changes in already identified language phenomena.

Abstract Monitor corpus; language use; trends; Slovene; neologisms; lexicography; newsfeed

1. Introduction

One of the challenges of lexicographers has always been staying on top of changes in a language. This has become even more crucial in recent years as the technological progress and the shift of dictionaries to the online media has raised the expectations of the users; as research shows (e. g. Kosem et al. 2019) uptodatedness is among the highest valued dictionary features. As a result, new words and senses have started to enter dictionaries at a much faster rate than ever before. The field of neology has attracted more and more attention, with the recent COVID-19 pandemic with all its new vocabulary being a good case in point.

In order to be able to obtain the information on new words and uses of words, one needs a monitor corpus. The main characteristic of monitor corpora is that new (recently published) texts are added on a regular basis, thus enabling monitoring language change. This is also one of the main challenges of monitor corpus compilation, namely being able to regularly obtain, annotate and upload the texts in the shortest time span possible. A monitor corpus for Slovene has long been called for by Slovenian lexicographers, linguists, and other users in need of information about current language use. This has become possible with the introduction of the JSI (Jozef Stefan Institute) Newsfeed service (Trampuš/Novak 2012), which collects news articles from websites across the world, covering over 35 languages. The service has already been used in various projects, including in the compilation of the Gigafida reference corpus of Slovene, version 2 (Krek et al. 2019).[1]

In September 2021, we started a new project called SLED (Monitor Corpus for Slovene and related language resources), which is funded by the Ministry of Culture of the Republic of Slovenia. The project has three aims, the main one being the development of the monitor corpus for Slovene, including the methodology for its regular updating. The second project aim is to regularly provide statistical datasets that will include information of interest to the wider public, for example trending words, new words, words with decreased usage, etc. The third project aim is to develop a tool for topic modelling that can be used on Slovene texts; in this way, we want to provide some topic categorization for each text included in the Monitor Corpus for Slovene, thus enabling more detailed analyses.

[1] https://viri.cjvt.si/gigafida/.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

In this paper, we first make an overview of related projects for languages other than Slovene, and then focus on various aspects of the first monitor corpus of (written) Slovene, called Trendi, including the methodology behind its compilation. We point out some of the important decisions that had to be made during corpus conceptualization. We also present the results of a user survey, which was used to get a better understanding of user needs and expectations. We demonstrate some of the ways in which the corpus will be integrated into the dictionary-making workflows in Slovenia. We conclude by presenting future plans related to the monitor corpus of Slovene, and related resources.

2. Related work

The concept of a monitor corpus is far from new. One of the first monitor corpora was the Bank of English,[2] which was first published in 1991. It contains over 650 million words and was used in the compilation of the COBUILD dictionary. Today, it is still a representative subset of the 4.5-billion-word COBUILD corpus. There is no information on when the corpus was last updated. Access to the corpus is very limited, with only the staff and students at the University of Birmingham having access.

Another influential corpus for English, in this case American English, is the Corpus of Contemporary American English (COCA; Davies 2008-),[3] which covers the period from 1990 onwards and contains over 1 billion words. It is a genre-balanced corpus, containing texts from eight different genres (spoken, fiction, popular magazines, newspapers, academic texts, TV and movies subtitles, blogs, and other web pages). The corpus was last updated in March 2020, which somewhat limits its "monitor" status.

Also part of the corpora at English-corpus.org is a regularly updated monitor corpus NOW (News on the web; Davies 2016–),[4] which contains nearly 15 billion words from web-based newspapers and magazines from 2010 to "yesterday" (if we borrow the wording from Mark Davies). As it is mentioned on the website, the corpus grows by about 180-200 million words per month.

Part of the same family as NOW and COCA is a more specialized Coronavirus corpus (Davies 2019-),[5] which spans the period from January 2020 to yesterday and contains over 1.4 billion words. Limited to the genre of web news in English, it grows about 3-4 million words per day.

There are also corpus resources for monitoring languages other than English, for example Timestamped JSI web corpora, which are available in 18 different languages and contain news articles collected by the JSI Newsfeed service. The corpora are available in the Sketch Engine corpus tool (Kilgarriff et al. 2004) and in addition to the usual Sketch Engine functions, the users can also use Trends (Herman 2013), a feature focused on identifying trends in word usage. The corpora contain texts from 2014 to April 2021 (time of the last update) and are of different sizes, with the English corpus containing approx. 60 billion words.

Similarly multilingual is the Google Books Ngram Viewer, which offers searching and various visualizations of word/ngram use over time (1500 to 2019). The resource could loosely

[2] https://cqpweb.bham.ac.uk/.

[3] https://www.english-corpora.org/coca/.

[4] https://www.english-corpora.org/now/.

[5] https://www.english-corpora.org/corona/.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

be called a corpus, as the data are based on texts and the user is able to get to the parts of the documents; however, the usual functions for linguistic investigation of corpora such as concordancers, collocations, etc. are not available.

There are many other monitor corpora in existence, which are used by lexicographic institutions and available only internally. An example of such a resource is ONLINE, a dynamic monitor of Czech, compiled by the Czech National Corpus. It contains approx. 6.3 billion words, coming from web news, discussions (under news articles), forums, and social networks (Facebook, Twitter, Instagram). The ONLINE corpus is in fact divided into two complementary corpora – ONLINE_NOW and ONLINE_ARCHIVE. ONLINE-NOW, which is updated daily, covers the period of the current month + the last six months, whereas the ONLINE_ARCHIVE covers the preceding period back to February 2017. At the beginning of each month, the contents of the oldest month of the ONLINE_NOW corpus are moved to ONLINE_ARCHIVE.

Until now, there were no monitor corpora of Slovene in existence. Nonetheless, recently, a resource called Language Monitor (Kosem et al. 2021) has been developed, which indicates trending words and N-grams in recent periods and is updated on a regular basis. Like Timestamped JSI web corpora, the Language Monitor also uses the IJS Newsfeed service to export news articles. After linguistic annotation (tokenization, lemmatization, morphosyntactic annotation, parsing) of texts, word lists are generated, and statistical calculations are conducted. This basically means that whenever Language Monitor is being updated, a sort of temporary monitor corpus is being created and a considerable manual effort is needed. Moreover, word lists provided are not linked to examples of use (e. g. corpus concordance), limiting their usefulness.

3. Trendi – a monitor corpus of Slovene

Services such as Language Monitor, which offer already prepared statistics for users, are more suitable for the general public; while lexicographers, linguists, and other language experts may find some of these options useful, they also need direct access to corpus data for their analyses. This is the motivation behind the SLED project, in which the first monitor corpus of Slovene, called Trendi, will be compiled.

3.1 Methodology

One of the main decisions in preparing Trendi was determining the time period covered by the corpus, and the regularity of its updates. As we learned from analysing a selection of monitor corpora of other languages, there was no uniform approach used. Our main principle was for Trendi to fill the gap not covered by the most recent version of the Slovene reference corpus, i. e. the Gigafida corpus, version 2.0 (Krek et al. 2019). Thus, with Gigafida 2.0 (1.13 billion words) covering the period from 1991 to 2018, the first version of Trendi covers the period from 2019 onwards. There are plans to publish Gigafida 3.0 towards the end of 2022, and to then make much more regular updates to the corpus, which will result in the monitor corpus covering a shorter period, and also being smaller in size.

Maintaining a close compatibility with the Gigafida corpus also means that the Trendi corpus covers (or monitors) the standard written Slovene language. The decision was based mainly on the needs of potential users of the corpus (translators, linguists, researchers,

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

computational linguists, etc.) but also on the fact that non-standard Slovene is being covered by other projects such as JANES (Jezikoslovna analiza nestandardne slovenščine, 'Linguistic Analysis of Nonstandard Slovene'; Fišer/Ljubesic/Erjavec 2018).

As far as updates of the Trendi corpus are concerned, a new version will be released every month, uploaded both to the CLARIN repository and the relevant concordancers.

3.2 Contents

All the contents of the Trendi corpus are at the moment obtained by using the IJS Newsfeed service (Trampuš/Novak 2012). The Newsfeed has already been used for linguistic projects like Language Monitor (Kosem et al. 2021), which indicates trending words and N-grams in recent periods and is updated on a monthly basis. The selection of newsfeed sources for Language Monitor was very inclusive, taking all Slovenian sources with at least 10 articles per year.

In the selection of the sources for the Trendi corpus, we wanted to be more rigorous. Also, we had to consider the fact that we wanted Trendi to represent standard written Slovene. For this reason, we joined forces with the compilers of the Gigafida corpus. We made a list of all Slovenian sources that were part of the newsfeed since 2019 and made an analysis of their contents. The initial list included 243 sources. 90 sources were immediately excluded because they were mostly foreign websites o websites with non-Slovenian content. A further 34 sources were excluded for various reasons: not being a news source (e.g. blogs, government and company websites), not covering standard Slovene (e.g. repositories of academic publications such as diplomas and theses), and being an aggregator of news from news sources which were already on the list. The final list included 110 sources, with the top 15 and the number of news items from 2019 to 2021 shown in Table 1.

Source	Number of articles
sta.si	260,080
rtvslo.si	97,924
siol.net	69,471
delo.si	65,415
24ur.com	61,623
dnevnik.si	47,749
vecer.com	45,548
novice.svet24.si	42,049
vestnik.si	41,525
zurnal24.si	39,220
ekipa.svet24.si	35,326
demokracija.si	26,604
gorenjskiglas.si	22,883
nova24tv.si	20,153
slovenskenovice.si	18,622

Table 1: Top 15 news sources by a number of articles (2019–2021) in the Trendi corpus

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

One thing to note is that some of the sources, e. g. sta.si, delo.si and dnevnik.si, have some of their content available only through subscription. As a result, such news items collected from their websites contain only a title, sometimes a subheading, and the first paragraph. This issue has been resolved by forming a close collaboration with the Gigafida corpus team, as they are in the process of signing contracts with source providers to send them the full contents on a regular basis. Once this procedure is established, the contents of Trendi (as well as Gigafida of course) will become even richer.

At the time of writing this paper, the first version of the corpus was being prepared, with the intention of including the data up to May 2022, so we did not yet have the exact details on its size. We have already made some preliminary calculations for 2019–2021 data, and the 2019 subcorpus contains nearly 12,5 million words per month, the 2020 subcorpus nearly 15 million words per month, and the 2021 subcorpus nearly 21 million words per month. One of the reasons for the continuous increase in size per year is the regular appearance of new websites, for example necenzurirano.si was launched in 2020 and is already 28th on the list of sources (per number of news items) with 8,494 news items. This finding also underlines the importance of continuously monitoring the Slovenian web space for new websites, and adding the relevant websites to the Trendi corpus.

3.3 Article collection and annotation pipeline

The texts for the Trendi corpus are being downloaded on a daily basis, in the JSON format. All the articles from each individual source are merged into a single daily file before annotation. The deduplication check, i. e. ensuring (via URL) that the same article is not downloaded more than once, is already performed by the JSI Newsfeed service. No further deduplication is conducted at the moment, although we are aware that very similar articles can be found in various sources, especially media ones. This is because we want to make it possible for the users of the corpus to analyse the contents of individual sources, compare two or more sources, etc. A different approach will probably be taken for the Gigafida corpus where the deduplication is done on a paragraph level (Krek et al. 2019). Such a step for a reference corpus is very much needed, also considering the fact that STA (sta.si) is a service for the distribution of original press releases, which means that many media websites prepare articles based on these pieces of information and often use a considerable portion of the contents.

During the annotation of the files, the processes of tokenization, lemmatization, morphosyntactic tagging, dependency parsing, and named entity recognition are performed. The annotation output, provided in the CONNL-U format, is converted into the TEI format, the format needed for the calculations of various statistics, and for the conversion into the VERT format, used by the KonText and NoSketchEngine concordancing tools.

At the moment we are still getting the data from the JSI Newsfeed, therefore the TEI files need to be put through an additional step of source filtering, using the list of 110 sources as described in section 3.2. In the near future, we intend to limit the newsfeed extraction to only the sources selected for the Trendi corpus (and relatedly Gigafida).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

3.4 Accessibility

The Trendi corpus will be accessible via two concordancers: KonText CLARIN.SI (https://www.clarin.si/kontext), originally developed for the Czech National Corpus (Machálek 2020), and NoSketchEngine CLARIN.SI (https://www.clarin.si/noske/). The concordancers are somewhat complementary: they share many features, but KonText offers the option of registration and with that saving of searches and favourite corpora, whereas NoSketchEngine offers certain additional features such as Keyword extraction.

The Trendi corpus will also be uploaded to the CLARIN.SI repository, in both CONNL-U and TEI formats. Normally, corpora are provided only in the TEI format, however, our computational team has advised us to include CONNL-U as well, as this format is often preferred for processing tasks. The corpus will be made available under the Creative Commons - Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0) license. Due to copyright restrictions, we intend to publish paragraph samples from each text, the procedure already used for other corpora of Slovene such as the ccGigafida 1.0 corpus (Logar et al. 2013). We are also considering providing the full version of the corpus to individual researchers, under the condition of a signed agreement.

3.5 User survey

In addition to providing access to the Trendi corpus, the SLED project also aims to provide users interested in most current language use with various statistics and similar information on word usage. In order to get a good understanding of user needs and preferences regarding current language trends, we conducted a survey. The survey was prepared on the 1KA platform (http://www.1ka.si/) and included eight questions; four questions were content-related, and four questions collected information on the respondents (gender, age, occupation, and field(s) of activity).

The survey was completed by 100 respondents, 82 females and 18 males. The majority of the respondents were between 26–55 years old, with groups of 26–35 (33%) and 46–55 (32%) having the highest shares. 61% of the respondents were employed in the public sector (e. g. education, health organizations, public administration), and 20% were self-employed. Other groups such as company employees (6), retired persons (4), unemployed (5), and students (3) were much smaller. In terms of the field of activity, where the respondents could choose more than one option, the following groups dominated: proofreading (60%), translating (46%), language lover (38%), academic research writing (34%), language research (32%), and creative writing and blogs (22%). 40% in total is represented by the respondents from various categories of language education (Slovene as L1 in elementary or secondary school, Slovene as L2, language subjects at a university level).

In the first question, the respondents were provided with six different scenarios[6] and they had to express their level of interest (not interested at all, not interested, neither interested nor interested, interested, very interested, don't know). As Table 2 shows, they were interested in all the scenarios, with "interested" and "very interested" covering between 74–88% of responses. The highest interest was expressed for the last scenario where the trends in usage for two or more words or word combinations could be compared. Also high was the

[6] An example of a particular scenario was provided for clarity.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

level of interest in the information whether the usage of a word or word combination is increasing or decreasing recently.

On the question of whether the information on current language use would be helpful for their work, over three quarters (76%) of the respondents replied with Yes, with only 9% answering No, and 15% opting for "I don't know".

The third question asked the respondents about the importance of various visualizations of language data: diagrams, tables, and word lists. The answers indicate that diagrams, tables, and word lists are all considered important for the respondents, with up to 87% (for word lists) of the respondents considering them important or very important. A closer inspection of the results reveals that the respondents tended to slightly prefer a more simplistic presentation of language data, with tables with figures attributed the lowest combined importance (64%).

Scenario	**Not interested at all**	**Not interested**	**Neither**	**Interested**	**Very interested**	**Don't know**
Words or word combinations typical of a certain period compared to another period (e.g. which words are much more frequent in February 2020 compared to February 2021)	2%	8%	12%	**48%**	30%	0%
The period in which a certain word or word combination is the most frequent (e.g. was the word "tycoon" really the most frequent word in the period 2008–2009?)	5%	6%	14%	**44%**	30%	1%
Is the use of a word or word combination increasing or decreasing? (e.g. is the use of the word "epidemic" on the rise or is it decreasing)	2%	5%	5%	42%	**46%**	0%
In which texts (by topic) is a word or word combination more frequent? (e.g. is the noun "forward" really most frequent in sports texts)	1%	5%	9%	37%	**46%**	2%
What is the category distribution of the use of a word or a word combination? (e.g. is the word combination "collective immunity" found only in medical texts or not?)	2%	4%	12%	**51%**	30%	1%
Which of two (or more) words or word combinations is used more frequently in recent years/months? (e.g. which of the words "anti-vaxxer" or "countervaxxer" is used more frequently'?)	1%	5%	6%	34%	**53%**	1%

Table 2: Answers to six selected scenarios

The last question was an open-ended one and offered the respondents an opportunity to provide their own suggestions or scenarios for providing information on current language trends. The suggestions can be grouped into the following categories:

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

- linkability or integration with other language resources and data access via API
- comparison of synonyms or related words (foreign words or loanwords vs Slovene equivalents)
- inclusion of examples of word usage, e. g. via links to corpus concordances
- monitoring different senses of words over time
- monitoring syntactic behaviour of words over time
- monitor multiword units (e. g. phrases) over time

Although the survey confirmed several objectives of the SLED project, the findings also made it clear that the community needs both online access to the Trendi corpus, as well as an online tool that facilitates the analyses of language trends beyond the scope of the Trendi corpus, and offers simple visualizations of complex statistical data.

Given that the project only promised statistics regularly uploaded to the CLARIN repository, we had to rethink the approach and have started preparing an online service that will be linked to various corpora, including Trendi, and will offer the users different options of analysing language data and exporting the results. The service, planned to be completed by the second half of 2022, will get the data via API from a data warehouse where we will store statistical information on word forms, lemmas, collocations, and other linguistic phenomena. The statistical information is calculated using the pipeline extension based on the corpus extraction tool LIST (Krsnik et al. 2019).

4. Integration of Trendi into the lexicographic workflow

Over the past year, while working on the Language Monitor and later on the conceptualization of the monitor corpus, we have already started working on the infrastructure that would support the needs of lexicographers. Of course, this goes beyond identifying neologisms, which is indeed the most common use of monitor corpora; lexicographers also need to identify and monitor potential future neologisms (words with a frequency below the threshold for inclusion into a dictionary), identify new uses of existing meanings (e. g. via collocations), and identify meanings, collocations and other phenomena which are already included in dictionaries and are used less and less frequently or not at all. At the moment, Slovenian lexicographers are very much hindered by the fact that they do not have direct access to language data beyond 2018 - this makes any language description immediately, at least to a certain extent, outdated.

One of the important pieces of the planned infrastructure is the data warehouse mentioned in section 3.5, which will serve as a repository of all possible information from corpora. The data warehouse will be linked with corpora and dictionary tools, and indirectly (after the lexicographers analyse the data) with the Digital Database for Slovene, which is being developed at the Centre for Language Resources and Technologies at the University of Ljubljana. Most importantly, the data warehouse will contain information that is at the moment not relevant or not yet relevant – for example, potential neologisms, which are at the moment not yet frequent enough or limited to too few sources, can be saved there (but not in the Digital Database). Furthermore, all identified bad collocation candidates from automatic extractions can be recorded there to avoid duplication of work in the future – based on our experience, getting rid of repeated inspection of bad data could save a considerable amount of time, especially in this day and age when corpora are very large.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

5. Conclusions and future plans

The Trendi monitor corpus for Slovene described in this paper is a very much needed resource for tracking trends in Slovene language use. Because of the various purposes, the corpus will be used for, it was paramount to prepare a sound and sustainable methodology for text collection and annotation, as well as for linguistic data extraction. This will facilitate lexicographic, linguistic, and other analyses, thus benefitting end-users of dictionaries and similar resources. In addition, the fact that Trendi will complement the Gigafida reference corpus will mean that there will now be corpus data on the Slovenian language from 1991 to yesterday.

More challenging tasks lie ahead. Among them is ensuring regular updates to the corpus, by which we mean both uploading new versions to the concordancers, but also identifying and adding new web sources. A detailed evaluation and possible improvement of the article collection procedure will be made and will include selecting a sample of articles from each ***source and identifying potential issues such as unwanted content (e.g. menus) being included,*** only part of the article being collected, etc.

Finally, the activity that is currently underway and which will improve Trendi, but also other corpora, even more, is the development of an automatic text categorization program. At the time of writing, we have been finalizing the list of text categories (e. g. politics, sport) and preparing the training corpora. In the coming month, the algorithm using supervised training will be developed and then tested on newly acquired articles, and more importantly on the Gigafida corpus. This development means that in the future lexicographers could also be provided with the category dispersion of different language phenomena.

References

Davies, M. (2008-): The Corpus of Contemporary American English (COCA). https://www.english-corpora.org/coca/ (last access: 23-03-2022).

Davies, M. (2016-): Corpus of News on the Web (NOW). https://www.english-corpora.org/now/ (last access: 23-03-2022).

Davies, M. (2019-): The Coronavirus Corpus. https://www.english-corpora.org/corona/ (last access: 23-03-2022).

Fišer, D./Ljubešić, N./Erjavec, T. (2018): The Janes project: language resources and tools for Slovene user generated content. Language Resources & Evaluation. https://doi.org/10.1007/s10579-018-9425-z.

Herman, O. (2013): Automatic methods for detection of word usage in time. Bachelor thesis. Masaryk University.

Kilgarriff, A./Rychlý, P./Smrz, P./Tugwell, D. (2004): The Sketch Engine. In: Williams, G./Vessier, S. (eds.): Proceedings of the Eleventh EURALEX International Congress, Lorient, France. Lorient, pp. 105–116.

Kosem, I./Krek, S./Gantar, P./Arhar Holdt, Š./Čibej, J. (2021): Language monitor: tracking the use of words in contemporary Slovene. In: Kosem, I./Cukr, M./Jakubíček, M./Kallas, J./Krek, S./Tiberius, C. (eds.): Electronic Lexicography in the 21st Century. Proceedings of the eLex 2021 conference. 5–7 July 2021, virtual. Brno, pp. 514-527. https://elex.link/elex2021/wp-content/uploads/2021/08/eLex_2021_33_pp514-528.pdf.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Kosem, I. et al. (2019): The image of the monolingual dictionary across Europe: results of the European survey of dictionary use and culture. In: International Journal of Lexicography 32 (1), pp. 92–114. https://doi.org/10.1093/ijl/ecy022.

Krek, S. et al. (2019): Corpus of Written Standard Slovene Gigafida 2.0, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1320.

Krek, S./Arhar Holdt, Š./Erjavec, T./Čibej, J./Repar, A./Gantar, P./Ljubešić, N./Kosem, I./Dobrovoljc, K. (2020): Gigafida 2.0: the reference corpus of written standard Slovene. In: Calzolari, N. (ed.): LREC 2020: Twelfth International Conference on Language Resources and Evaluation: May 11–16, 2020, Marseille, France. Paris: ELRA – European Language Resources Association. 2020, pp. 3340–3345. http://www.lrec-conf.org/proceedings/lrec2020/LREC-2020.pdf.

Krsnik, L. et al. (2019): Corpus extraction tool LIST 1.2, Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1276.

Logar, N./Erjavec, T./Krek, S./Grčar, M./Holozan, P. (2013): Written corpus ccGigafida 1.0, Slovenian language resource repository CLARIN.SI, ISSN 2820-4042. http://hdl.handle.net/11356/1035.

Machálek, T. (2020): KonText: Advanced and Flexible Corpus Query Interface. In: Proceedings of LREC 2020, pp. 7005–7010.

Trampuš, M./Novak, B. (2012): The internals of an aggregated web news feed. In: Proceedings of 15th Multiconference on Information Society 2012 (IS-2012). http://ailab.ijs.si/dunja/SiKDD2012/Papers/Trampus_Newsfeed.pdf.

Contact information

Iztok Kosem
Jožef Stefan Institute & Faculty of Arts, University of Ljubljana
iztok.kosem@ijs.si

Acknowledgements

The project *Spremljevalni korpus in spremljajoči podatkovni viri* (SLED) is funded by the Ministry of Culture of the Republic of Slovenia.

The authors acknowledge the financial support from the Slovenian Research Agency (research core funding No. P6-0411, *Language Resources and Technologies for Slovene*)

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Simon Krek/Polona Gantar/Iztok Kosem

EXTRACTION OF COLLOCATIONS FROM THE GIGAFIDA 2.1 CORPUS OF SLOVENE

Abstract This paper describes a method for extracting collocation data from text corpora based on a formal definition of syntactic structures, which takes into account not only the POS-tagging level of annotation but also syntactic parsing (syntactic treebank model) and introduces the possibility of controlling the canonical form of extracted collocations based on statistical data on forms with different properties in the corpus. Specifically, we describe the results of extraction from the syntactically tagged Gigafida 2.1 corpus. Using the new method, 4,002,918 collocation candidates in 81 syntactic structures were extracted. We evaluate the extracted data sample in more detail, mainly in relation to properties that affect the extraction of canonical forms: definiteness in adjectival collocations, grammatical number in noun collocations, comparison in adjectival and adverbial collocations, and letter case (uppercase and lowercase) in canonical forms. The conclusion highlights the potential of the methodology used for the grammatical description of collocation and phrasal syntax and the possibilities for improving the model in the process of compilation of a digital dictionary database for Slovene.

Keywords Collocations; discovering collocations in corpora; digital collocation database

1. Introduction

Large text corpora and tools for their processing created over the last three decades have enabled the development of various methods for the automatic extraction of multi-word units from corpora, mainly for the purpose of compilation of dictionary resources, for natural language processing tools, and for the development of various language applications.[1]

Multi-word expression extraction procedures typically exploit a mechanism that recognises sequences of lexical units on the basis of their morpho-syntactic annotation in the corpus and statistical measures that determine co-occurrence values. The most recognized and established model, especially in the field of lexicography, is the word sketch model in Sketch Engine, which operates on the basis of a word sketch grammar (Krek/Kilgarriff 2006; Krek 2015; Gantar 2015; Kosem et al. 2018) and a lemmatized and POS-tagged corpus.[2] Our aim, however, was to devise a methodology for extracting collocation data from the Gigafida corpus that upgrades the existing system and is based on the assumption that collocation candidates can be successfully extracted from a syntactically parsed corpus, using labelled dependency relations and morphological features of the heads and the dependents in the dependency tree.

In this paper we describe a methodology for automatic extraction of collocations from the Gigafida 2.1 corpus, based on definitions of syntactic relations within a phrase, also taking into account some statistical parameters. First, we present the extraction procedure and the

[1] elexiFinder web service yields 423 results in six languages for the search "collocation": http://finder.elex.is/intelligence?conditions=3-wikidata:Q1122269-collocation&percentile=100&dataType=news&dataType=video&tab=items&type=articles&articlesSortBy=date (last access: 25-03-2022).

[2] Sketch grammar for Slovene (Krek in Kilgarriff 2006) was used in the Communication in Slovene project (Krek 2015) in the creation of the Slovene Lexical Database (Gantar 2015) and the Collocation Dictionary of Modern Slovene (Kosem et al. 2018).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

database of extracted collocations (Krek et al. 2021). We evaluate the extracted data based on quantitative and qualitative linguistic analyses. In conclusion, we highlight the potential of the methodology and the resulting open data for the grammatical description of collocations and phrase syntax, as well as the possibilities for improving the model in the construction of a digital dictionary database for Slovene.

2. Automatic extraction of collocations from the corpus

In this section we describe the formal description of collocation structures in an XML file (2.1), which is the core part of the new collocation extraction methodology. The most important part of the description is included in the definition of collocation structures (2.2), which consists of a description of the components of a collocation, the syntactic relations between them, and the various constraints on a) the identification of the components in the corpus, and b) the extraction of the final canonical forms of the collocation. In the last part of the section (2.3), we describe the automatic extraction procedure of collocations from the corpus, based on the proposed system.

2.1 Formal description of collocations

For the purposes of the new methodology, it was necessary to define more precisely the term "collocation", which is described in Gantar/Krek/Kisnik (2021). In defining the morpho-syntactic structure, we started from the previously defined grammatical relations in the Word Sketch tool for Slovene (Krek 2015). Starting with the POS-tagging annotation level, we added a syntactic parsing level, where we defined dependency syntactic relations within a collocation. Statistical and frequency data were considered both at the level of the lemma and the collocation as a whole, which has been shown to be an appropriate procedure in previous automatic extractions of collocations from the corpus (Gantar/Kosem/Krek 2016). At the same time, frequency data were also taken into account when determining the "representation" or the output form of the extracted collocation, i. e. the form in which the collocation should be included in the dictionary. In the process of creating a new formalism for collocation extraction, most of the collocation structures included in the Slovene Lexical Database were translated from the Sketch Engine grammar into the new formalism. The new method differs from the existing Sketch Engine methodology (Kilgarriff et al. 2014) in the following aspects:

- instead of the Corpus Query Language (CQL) used in the Sketch Engine, which mainly takes into account POS-tagging annotation, the new method uses its own system to define constraints on any level of annotation, from morphology (parts-of-speech and their properties), syntactic dependency relations, concrete lexical items, and any other types of annotation that can be used for other purposes, e. g. semantic roles, semantic types, word senses, etc.;
- in the new system, verbal structures are explicitly separated in terms of negation (expressed by a negation particle or by a verb) and reflexiveness (expressed by a free verb morpheme or a reflexive pronoun);
- unlike in the Sketch Engine system, identification numbers and (syntactic structure) names do not differ according to whether the starting point is the first or the second collocator in the collocation;

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

- the human-readable syntactic structure names directly reflect the characteristics of the individual collocation components in terms of their parts-of-speech and grammatical properties, according to the Multext-East/JOS annotation system;
- in addition to the constraints (restrictions) enabled by the CQL system, it is also possible to specify which of the forms of each component found in the corpus should be used in a specific collocation, according to the options offered within the pre-defined canonical collocation form in a specific structure (representation).

The total number of collocation structures in the DDD system is (currently) 82, of which, counting by collocator pair, there are six that include negation, 25 with reflexive verb structures, and 26 combinations with prepositions. Like in the Sketch Engine, collocators belong to four (content) word types: nouns, verbs, adjectives, and adverbs.

For human-readable codes, we use a short combination of the included morphosyntactic categories and features according to the MTE/JOS system (Erjavec et al. 2010, 2011). This is important for linguistic use, while the identification number is important for computational use.

In Table 1 below, we provide an example of a selection of ten structures, the first five according to the number of collocations extracted, the remaining five for the purpose of displaying the tags in all nine columns/categories.

ID	CODE	EXAMPLE	TRANSLATION	1	2	3	4	5	6	7	8	9	COL No.
34	p0-s0	svetovno prvenstvo	world championship		p0						s0		720,605
53	s0-s2	direktor podjetja	company director		s0						s2		518,199
70	s0-gg	raziskava pokaže	research shows		s0						gg		385,018
23	gg-s4	podpisati pogodbo	to sign a contract		gg						s4		270,965
15	gg-d-s5	imeti v mislih	to have in mind		gg			d			s5		235,771
30	p0-vp-p0	domač in tuj	domestic and foreign		p0		vp				p0		32,127
77	s1-gp-s1	nogomet je šport	football is a sport		s1					gp	s1		26,520
72	s0-l-gg	trditev ne drži	the claim is not true		s0				l		gg		19,400
95	l-gg-zp-ggn	ne uspeti se uvrstiti	fail to qualify	l	gg	zp					ggn		479
94	gg-zp-ggn-zp	odločiti se vrniti se	decide to return		gg	zp					ggn	zp	5

Table 1: Collocation structures according to the categories represented and the number of cases extracted

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

To create an algorithm for the automatic extraction of collocations from the corpus, we have created a formalism that includes all the necessary information in XML format. This allows for later adaptation, addition, or reduction of structures in further extractions processes. We describe the formalism in more detail below.

2.2 Definition of syntactic structures

The individual syntactic structure is defined by the `<syntactic_structure>` element, which provides three mandatory attributes. These include:

- structure ID: `@id`,
- a human-readable code of the structure: `@label`,
- structure type: `@type`.[3]

The structure definition relies on specific tag sets and corpus tagging systems, so at the first level under structure in the `<system>` element we define the tagging system which we will follow. This contains the `@type` attribute, whose value defines the selected tagging system. In the context of the project where this procedure was developed, we applied the JOS or Multext-East tagging system to the morphosyntactic and syntactic tagging of the Gigafida 2.1 corpus, both at the morphosyntactic and syntactic levels.[4]

Within the specific labelling system, we further define three distinct groups of information:

- the individual words or elements that make up a collocation – the components,
- links between elements at the syntactic level – dependency tree,
- constraints and other information needed to extract collocations – structure definition.

2.2.1 Components

The components are defined in the `<components>` element containing the `@order` attribute. This may contain the values "fixed" and "variable". This attribute specifies whether the automatic parsing of the structure and the extraction of the components takes into account the order of the components as specified in the structure definition, or whether the predominant order as it is found in the corpus is taken into account – i. e. the order of the components of a particular collocation that represents the majority of the sentences in the corpus. An example of a structure where the sequence is variable is the adverb-verb phrase r-gg (ID 43), where the output of the collocation will vary according to the typical occurrence of the two elements or the adverb semantic group, e. g. *ostati doma* /to stay at home/ (gg-r) vs. *veliko pomeniti* /to mean a lot/ (r-gg).

All components are listed in the `<component>` (sub)elements, containing several attributes:

- the component identification number: `@cid`,
- the component human-readable code: `@label`,

[3] In this paper we consider 82 structures belonging to the type="collocation". The other types are: type="single" for single-word lexemes and type="other" for multi-word units.

[4] New grammar of contemporary standard Slovene: sources and methods.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

- component type: `@type`,
- component status: `@status`.

The `@label` attribute repeats information from the entire structure code, but only references the part that defines the specific component. The `@type` attribute defines the core of the components and can contain two values: "core" and "other". Core components are the actual components included in the collocation structure, which are defined in the collocation code and are also included in its output. Components marked with "other" are used in cases where, in order to correctly identify a collocation in a specific structure (in a corpus sentence), we need to define additional elements which are either mandatory or prohibited. Components defined with the value "other" in the `@type` attribute must therefore also contain the `@status` attribute, where the values "obligatory" and "forbidden" are allowed. The former specifies that the component must necessarily be in the sentence in which the collocation is found, even if the component itself is not included in the output as part of the collocation. The second value has the reverse role: a component with the status value "forbidden" defined in the structure must not be present in the corpus sentence.

2.2.2 Syntactic relations

The next major unit of structure description is the `<dependencies>` element, which defines the syntactic relations between components. Three attributes (`@from`, `@to`, `@label`) are mandatory in the `<dependency>` (sub)elements, the number of which must correspond to the number of components. An additional (optional) attribute `@order` is possible:

- origin of the dependency tree link (MTE/JOS): `@from`,
- the target of the dependency tree link (MTE/JOS): `@to`,
- link identifier (MTE/JOS): `@label`,
- order of the linked components: `@order`.

The last attribute `@order`, with allowed values "to-from", "from-to" or the default value "any", determines whether the two components associated with this dependency link must be in a specific word order in the sentence or not. In the case of the ID 34 structure given above, the use of the `@order` attribute means that the adjectival modifier must precede the nominal head in the corpus sentence, in order for the collocation to be recognised as corresponding to this structure. The hash character (#) used as a value in the `@from` and `@label` attributes denotes that we do not want to restrict the dependency head of this component, or its label. Therefore, it replaces any origin or label of the link.

2.2.3 Restrictions and output

The most extensive part of the formal description of the structure is the `<definition>` element, with the `<restriction>` element defining the constraints for each component when querying the corpus, and the `<representation>` element defining the variables of the extracted collocations. The latter contains only components that are defined as core and are actually included in the collocation output.

The `<restriction>` element contains a `@type` attribute that specifies at which annotation level the restriction information will be found. Currently, the values "morphology" and "lexis" are used. The first value specifies that the constraints will refer to the POS-tagging

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

annotation level in the corpus. The second value denotes that when identifying a component, we restrict ourselves to specific occurrences, either at the level of the word form or the lemma as found in the corpus.

The `<representation>` element defines variables in the output of the extracted collocations. These will also be found in the `<feature>` element, but with different attributes. The `@rendition` attribute defines the type of information to be used in the output. The values "lemma" and "word_form" specify that we will use either a lemma or one of the word forms of the component as found in the corpus. The value "lexis" in the `@rendition` attribute means that we will use an element that we may not have found in the corpus, but we want it in the output anyway, in place of the component from the collocation. To make this element concrete, we use the `@string` attribute with a chosen string of letters, which represents the actual output in the collocation. An example of such use is negation structures, where we want to control the output of the negation particle *ne* "not", even though the variant *ni* "not" would be more common in the corpus, or the negated forms of the verb *biti* (to be), e. g. *nisem* (I am not).

Furthermore, in the `<feature>` element, the `@selection` attribute (in combination with the `@rendition` attribute) determines which of the possible word forms found in the corpus should be chosen for the output in the collocation. The possible values in the `@selection` attribute are: "all", "msd", or "agreement". The first ("all") means that we include all forms of the component found in the corpus. This is useful, for example, in the case of reflexive pronouns, which have the possible forms *se* and *si* in different combinations, and if both are found in the corpus, both are also rendered in the collocation – *izogibati se/si pogovoru* (avoid the conversation).

The value "msd" in the `@selection` attribute is used in cases where we want to specify which of the forms found in the corpus is chosen for the output, according to its morphosyntactic properties. Individual properties in the same element are defined by combining the property and its value, e. g.

```
<feature selection="msd" case="nominative"/>
```

This means that we want the algorithm to output the (most common) nominative form of the word it found in the corpus.

The value "agreement" in the `@selection` attribute is used in case we want the extracted form of a component to agree in certain properties with the same properties defined in another component, which is defined in the `@msd` and `@head_cid` attributes. The first attribute defines the properties to be matched, the second one refers to the component ID containing the properties to be considered for matching. For example:

```
<feature selection="agreement" msd="gender+number+case"
head_cid="2"/>
```

The example specifies that the two components must agree in gender, case, and number.

The elements described above (in combination with the categories, properties and values in the chosen tagging system) define all 82 collocation structures that we used to extract a total of over 4 million collocations from the Gigafida 2.1 corpus, which we describe below.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

2.3 Extraction of collocation data from the Gigafida corpus 2.1

For the automatic extraction of collocation candidates, we used the Gigafida 2.0 corpus (Krek et al. 2020), published in 2018. The upgrades from the previous version include, among others, improvements in lemmatisation and POS-tagging, the removal of non-standard texts, and the inclusion of underrepresented and more recent texts. The Gigafida 2.1 version of the corpus, which was used for collocation extraction, also includes an additional level of syntactic parsing, semantic role labelling, and named entity recognition.

The final collocation database (Krek et al. 2021) contains 4,002,918 collocations, automatically extracted from the corpus, based on the definition of 82 collocation structures. The minimum frequency of extracted units in the database is 10. The output is divided by structure into 81 files in tabular format, with comma as a separator (CSV format). The number of files in the database is lower than the number of structures because structure ID-97 (l-gg-zp-ggn-zp, *ne bati se pokazati se* 'not to fear not to show') did not produce results with collocations above a frequency of 10. All collocations are assigned with the following information in 26 columns (Table 2).

Col	Column heading	Description
1	Structure_ID	structure identification number
2	C1_Lemma	lemma of the first component
3	C1_Representative_form	word form of the first component (according to the structure definition)
4	C1_RF_msd	morphosyntactic description for the word form of the first component
5	C1_RF_scenario	scenario for the word form output of the first component
6	C1_Distribution	number of different collocations containing the C1 component lemma (within the structure)
7	C1_lemma_structure_frequency	number of corpus sentences with collocations containing the C1 lemma (within the structure)
8	C2_Lemma	SAME INFORMATION FOR COMPONENTS C2/3/4/5
...	...	...
21	Colocation_ID	collocation identification number
22	Joint_representative_form_fixed	output of the canonical form of the collocation (according to the structure)
23	Joint_representative_form_variable	a list of the most frequent forms of collocation (according to word order)
24	Frequency	frequency of collocation
25	logDice_core	collocation strength calculation (logDice)
26	Distinct_forms	number of different forms of collocation

Table 2: Types of data in the collocation database for each collocation structure

In the following section, we describe some basic data about the extracted collocations and some of the more important advantages of the new method.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

3. Linguistic aspects of collocation database description

In the third section, we discuss selected linguistic topics of interest for the analysis of extracted collocations, including (in)definite forms in adjectival collocations (3.1), grammatical number (dual/plural vs. singular) in nominal collocations (3.2), degree (base vs. comparative and superlative) in adjectival and adverbial collocations (3.3), and uppercase vs. lowercase script (3.4).

For the purpose of linguistic evaluation, cumulative data for collocation candidates were extracted for 88 lemmas, with a minimum frequency of at least two occurrences. The number of collocations considered was, therefore, higher than the number of occurrences contained in the database for these lemmas, where the frequency limit is 10. Given the previous extraction methodology, it is mainly the representational part of the definition that is of interest for the evaluation, which is described in more detail below. The possibility to control the collocation output means that we can allow variability in the selected collocation elements, which reflects the actual situation in the corpus for specific collocation candidates. In the case of the 82 structures selected, the variability was allowed at the level of:

- definite (or indefinite) nominative forms of adjectives in the masculine singular
- grammatical number in collocations with nouns
- degree in collocations with adjectives and adverbs
- word forms in collocations written in lower or upper case.

The findings are described in more detail for these four categories (cf. Pori/Kosem 2021).

3.1 Definiteness in adjectival collocations

The new method makes it possible to highlight more adequately the relation between definite and indefinite forms of adjectives as they appear in real usage. We extracted the first 30 collocational candidates, ranked by logDice and filtered by:

- a morpho-syntactic description (the adjectival element must exhibit the following properties: masculine, singular, nominative);
- the difference between the attributed corpus lemma (which, according to lexicon convention, is always in the indefinite form, if it exists) and the extracted form of the adjective;
- a corpus frequency of at least 10 occurrences (the limit used in the collocation database;,
- the occurrence of each component in at least two collocations.

As an example: the indefinite form of the adjective *akuten* 'acute' (masculine, nominative, singular) is *akuten*, and the definite form is *akutni*.

As expected, the candidates with the definite form are often terms in a specific field, e.g. *etilni alkohol* 'ethyl alcohol', *akutni sindrom* 'acute syndrome', *avtomatični stabilizator* 'automatic stabiliser', *akutni hepatitis* 'acute hepatitis', etc. Similarly, they include names of animals and plants: *kodrasti pelikan* 'Dalmatian pelican', *kodrasti ohrovt* 'curly-leaf kale', *dolgoživi bor* 'Great Basin bristlecone pine', etc.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

The definite form of the adjective is also used to express a number of fixed phrases or expressions that are both in terminological use in a particular field and also part of the general vocabulary, e.g. *tuji jezik* 'foreign language', *letni dopust* 'summer holiday', *materni jezik* 'mother tongue', *solatni bife* 'salad buffet', *samopostrežni bife* 'self-service buffet', *kolektivni dopust* 'collective holiday', *neplačani dopust* 'unpaid leave', etc.

The method used in previous extractions (cf. Krek 2006) only allowed lemmas for adjectival elements as outputs in similar structures, which led to the export of "unnatural" collocations, e.g.: *tuj jezik* 'foreign language', *leten dopust* 'summer holiday', *solaten bife* 'salad buffet', etc., which was basically due to choices made by the creators of the tag set and the output of the POS-tagger for Slovene.

In addition to clear-cut choices, there are two cases where both indefinite and definite forms are acceptable, since the adjective can be understood as expressing either a species or a property: *bakren kotliček* 'copper kettle, *dobrodelen bazar* 'humanitarian bazaar'. However, even in these cases, the predominance of the definite form in the corpus data suggests that this form might be more suitable as a dictionary headword. It can be concluded that when it comes to the choice of adjectival (in)definite forms, allowing for variability produces the intended results.

3.2 Grammatical number in noun collocations

For noun components, most structures allow for variability in grammatical number. This means that the choice of the singular, dual or plural form of a noun is left to the observed corpus frequency, regardless of the grammatical case or other properties. We analysed the first 30 collocations from the set of 88 headwords where the plural form was extracted for (any) noun. We sorted them by logDice and filtered them by exhibiting the plural property of the noun, with the frequency of at least 10 and with at least three occurrences in the corpus.

Collocations that indicate phraseology are quickly noticeable, e.g. *briti* [*norce*] [*PL*] (*iz koga/česa*) 'to make a fool of (someone/something)', *brusiti* (*si*) [*kremplje*] [*PL*] 'to sharpen (one's) claws – to prepare for an (aggressive) action'. In principle, plural forms can be expected to be justified in these cases, but these units have a logic of their own and, in most cases, considerable variation can also be expected. The remainder can be divided into three categories – collocations where the plural form is a) justified or necessary, e.g. *drama s* [*talci*] [*PL*] 'hostage drama'; b) unjustified or incorrect, e.g. [*kotli*] [*PL*] *na biomaso* 'biomass boilers'; c) perhaps more common, but one would expect the dictionary form to be singular, e.g. *brinove* [*jagode*] [*PL*] 'juniper berries'. Categories a) and b) are correctly represented in most cases. The largest group is c), where one might expect the singular form to be more likely, but the plural form is neither wrong nor "annoying".

We also examined the extracted dual grammatical number forms on a slightly smaller set. In the 88 selected headwords, there are no eligible dual forms at the top of the collocation set (sorted by logDice). If we look at an extended set of extracted dual forms from the whole collocation database, it is possible to find cases where dual forms would be justified, especially in the case of paired (human-animal) organs or in similar pairing situations, e.g. [*ledvici*] [*DU*] *odpovesta* 'kidneys fail', *uiti med* [*nogama*] [*DU*] 'to escape between the legs', *enojajčni* [*dvojčici*] [*DU*] 'identical twins', etc. We can conclude that despite the predominance of the plural (or dual) form shown in the corpus, the existing criterion for the

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

extraction of the plural form (more than half) is mostly not justified. Statistical criteria for narrowing down the extraction of plural forms to category (a) from the above analysis remains a task for further work.

3.3 Degree in adjectival and adverbial collocations

In the case of adjectives and adverbs, variability is also checked at the level of degree – i. e. if the corpus in a particular collocation is dominated by the comparative and superlative forms, as opposed to the base form, which is also the default form of the lemma in adjectives and adverbs. Relatively few collocations were extracted for the 88 headwords that necessarily require the use of the comparative or superlative form. Most of these are related to adjectival base forms that are rarely used (e. g. *blizek* 'close') or where there is a marked semantic difference between the two forms. For example, *blizka bolnišnica* 'a nearby hospital', *ljuba kitara* 'favourite guitar'. In almost all the other cases, it would seem that the comparative and superlative forms would not be strictly wrong, but the output would be problematic if the collocation with the base form were ignored due to the majority of the two non-base forms. Analysis suggests that it would be more appropriate to consider comparative and superlative forms in the extraction only in cases where the base forms are not found in the corpus at all.

3.4 Upper or lower case

For all extracted components, we also allow for variation at the level of upper and/or lower case. This gives us insight into their dominant occurrence in the corpus and has provided interesting results. We analysed 30 of the most frequent collocations for 88 headwords, where one of the components (the dominant one) is written in uppercase or in capital letters. We sorted the collocations by absolute frequencies from the Gigafida 2.1 corpus and filtered them by the number of forms at least 3.

As expected, names of institutions, publications, and geographical names are dominant on the list, e. g. *ljubljanska Drama* 'name of a theatre from Ljubljana', *Slonokoščena obala* 'Ivory Coast – a country in Africa'. Understanding the use of upper or lower case is useful, particularly because it clearly indicates that the extracted collocation is not part of the general vocabulary; these are mainly proper names that we do not want to include in dictionary databases or the analysis of collocation data.

4. Conclusion

In this paper we describe a new procedure for extracting collocation candidates from a chosen corpus. The new formalism for collocation extraction takes into account various levels of corpus annotation, for which it uses its own (generic) system to define constraints at any level of annotation, ranging from POS-tagging and grammatical properties of word forms, syntactic (dependency) relations, concrete lexical items, and other levels of annotation, e. g. semantic roles, semantic types, etc. To automate the extraction process – in addition to constraints that take into account any annotation level in the corpus – the new system also allows us to specify which of the forms of each component found in the corpus should be included in a specific collocation, according to possibilities limited by the canonical collocation form in a specific collocation structure.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

In the second part of the article, we highlighted some of the variability in collocation extraction that the new system allows. This includes: the relationship between the definite and indefinite forms of the masculine singular nominative in adjectives; the singular, dual or plural forms of the nouns; the degree (comparative, superlative) of the adjective and adverb; the capitalisation of all elements of collocations. Our analysis shows that the possibility to manage the extracted forms is useful, but in most cases the threshold should be raised or the parameters further defined to take these phenomena into account when extracting collocations.

5. Future work

The main priorities for future consideration are:

1) Upgrading collocation structures from binary to extended collocations. In the existing 82 syntax structures, only binary collocations are considered. In some cases, it may be useful to include additional elements in collocations. While keeping the basic binary collocation, it would be beneficial to mention the additional element explicitly. For example: *govoriti jezik* --> *govoriti* [*angleški, francoski, ...*] *jezik* 'to speak a language --> to speak [English, French, ...] language'. The system is already set up in a way that allows existing structures to be combined into a more complex set that also takes into account the identification of extended collocations.

2) Taking into account statistics on distribution by corpus source or genre. It is possible to add various metadata from the corpus, such as textual distribution (the number of different texts in which a collocation appears) or distribution by source, to the statistics attributed to extracted collocations in the existing system. Similarly, the temporal dimension can be taken into account, meaning that we also take into account the distribution by year, which is not offered by current statistics.

3) More precise specification of the parameters for the form of the collocation output: as the analysis has shown, the possibility of managing the output of collocation forms is an important mechanism that helps to automatically extract more natural collocation forms. It is possible to build on the existing mechanism and create more precise specifications about when additional properties are taken into account and when not.

4) Consideration of other levels of annotation: semantic tagging of corpora (named entity recognition, semantic types, semantic frames, word sense disambiguation, wikification, etc.) has made significant progress, especially with the introduction of new technologies – deep neural networks. This means that future work should also take into account the next – semantic – level of annotation, which is likely to yield even better results, especially when considering clustering collocations and mapping them to corresponding dictionary senses.

References

Erjavec, T./Krek, S./Fišer, D./Ledinek, N. (2011): Project JOS: linguistic annotation of Slovene. Institut Jožef Stefan, Odsek za tehnologije znanja. http://nl.ijs.si/jos/ (last access: 23-03-2018).

Erjavec, T./Krek, S./Arhar, Š./Fišer, D./Ledinek, N./Saksida, A./Sivec, B./Trebar, B. (2010): Oblikoskladenjske specifikacije JOS V1.1 2010-03-07. http://nl.ijs.si/jos/msd/html-sl/index.html (last access: 23-03-2018).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Gantar, P. (2021): Zapis frazeoloških enot v Leksikonu večbesednih enot za slovenščino. In: Arhar Holdt, Š. (ed.): Nova slovnica sodobne standardne slovenščine: viri in metode. Ljubljana, pp. 198–230.

Gantar, P. (2015): Leksikografski opis slovenščine v digitalnem okolju. Znanstvena založba Filozofske fakultete. http://www.ff.uni-lj.si/Portals/0/Dokumenti/ZnanstvenaZalozba/e-knjige/Leksikografski.pdf (last access: 23-03-2018).

Gantar, P./Gorjanc, V. (2015): Obrazilo -en/-ni v slovarski obravnavi pridevnikov. In: Smolej, M. (ed.): Slovnica in slovar – aktualni jezikovni opis. Ljubljana, pp. 233–241.

Gantar, P./Kosem, I./Krek, S. (2016): Discovering automated lexicography: the case of Slovene lexical database. In: International Journal of Lexicography 29 (2), pp. 200–225.

Gantar, P./Krek, S./Kosem, I. (2021): Opredelitev kolokacij v digitalnih slovarskih virih za slovenščino. In: Kosem, I. (ed.): Kolokacije v slovenščini. Ljubljana, pp. 15–41.

Gantar, P./Krek, S./Krsnik, L. (2021): Strojno berljiv Vezljivostni leksikon slovenskih glago-lov. In: Arhar Holdt, Š. (ed.): Nova slovnica sodobne standardne slovenščine: viri in metode. Ljubljana, pp. 259–297.

Gorjanc, V./Gantar, P./Kosem, I./Krek, S. (eds.) (2015): Slovar sodobne slovenščine: problemi in rešitve. Ljubljana.

Kilgarriff, A./Baisa, V./Bušta, J./Jakubíček, M./Kovář, V./Michelfeit, J./Rychlý, P./Suchomel, V. (2014): The Sketch Engine: ten years on. In: Lexicography ASIALEX 1.

Kosem, I./Krek, S./Gantar, P./Arhar Holdt, Š./Čibej, J./Laskowski, C. (2018): Kolokacijski slovar sodobne slovenščine. In: Fišer, D./Pančur, A. (ed.): Zbornik konference Jezikovne tehnologije in digitalna humanistika. Proceedings of the Conference on Language Technologies & Digital Humanities, Ljubljana, September 20–21, 2018. Ljubljana, p. 133.

Kosem, I./Gantar, P./Krek, S./Arhar Holdt, Š./Čibej, J./Laskowski, C./Pori, E./Klemenc, B./Dobrovoljc, K./Gorjanc, V./Ljubešić, N. (2019): Collocations dictionary of modern Slo-vene KSSS 1.0. Ljubljana. https://www.clarin.si/repository/xmlui/handle/11356/1250 (last access: 23-03-2018).

Krek, S. (2015): Leksikografska orodja za slovenščino: slovnica besedlnih skic. In: Gorjanc, V./Gantar, P./Kosem, I./Krek, S. (ed.): Slovar sodobne slovenščine: problemi in rešitve. Ljubljana, pp. 358–378.

Krek, S./Kilgarriff, A. (2006): Slovene word sketches. In: Erjavec, T./Žganec Gros, J. (eds.): Language technologies. Proceedings of the 9th International Multiconference Information Society IS 2006, Ljubljana, 9–10 October 2006. Ljubljana: Institut "Jožef Stefan", p. 62.

Krek, S./Arhar Holdt, Š./Erjavec, T./Čibej, J./Repar, A./Gantar, P./Ljubešić, N./Kosem, I./Dobrovoljc, K. (2020): Gigafida 2.0: the reference corpus of written standard Slovene. In: Calzolari, N. (ed.): LREC 202: Twelfth International Conference on Language Resources and Evaluation, Marseille, May 11–16, 2020. Marseille, p. 3340.

Krek, S./Gantar, P./Kosem, I./Dobrovoljc, K./Arhar Holdt, Š./Čibej, J./Laskowski, C./Klemenc, B./Krsnik, L. (2021): Frequency lists of collocations from the Gigafida 2.1 corpus. Ljubljana. http://hdl.handle.net/11356/1415 (last access: 23-03-2018).

Pori, E./Kosem, I. (2021): Evalvacija avtomatskega luščenja kolokacijskih podatkov iz besednih skic v orodju Sketch Engine. In: Kosem, I. (ed.): Kolokacije v slovenščini. Ljubljana, pp. 43–77.

Ramisch, C. (2020): Computational phraseology discovery in corpora with the MWE-TOOLKIT. In: Corpas Pastor, G./Colson J-P. (ed.): Computational Phraseology. Amsterdam/Philadelphia, pp. 111–134.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Ramisch, C./Savary, A./Guillaume, B./Waszczuk, J./Candito, M./Vaidya, A./ Barbu Mititelu, V./ Bhatia, A./Iñurrieta, U./Giouli, V./Güngör, T./Jiang, M./Lichte, T./Liebeskind, Ch./Monti, J./Ramisch, R./Stymne, S./Walsh, A./Xu, H. (2020): Edition 1.2 of the PARSEME Shared Task on Semi-supervised Identification of Verbal Multiword Expressions. In: Markantonatou, S./McCrae, J./Mitrović, J./ Tiberius, C./Ramisch, C./Vaidya, A./Osenova, P./Savary, A. (eds.): Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons, Barcelona (Online), December 2020. Barcelona, p. 107.

Acknowledgements

This paper was written within the framework of the research project New Grammar of Modern Standard Slovene: sources and methods (J6-8256) and the programme groups Slovene Language – Basic, Contrastive and Applied Research (P6-0215) and Linguistic Resources and Technologies for the Slovene Language (P6-0411), funded by the Slovenian Research Agency.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/ Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Meike Meliss/Vanessa González Ribao

VERGLEICHBARE KORPORA FÜR MULTILINGUALE KONTRASTIVE STUDIEN

Herausforderungen und Desiderata

Abstract This contribution aims to show the necessity of working in the development of multilingual corpora and appropriate tools for multilingual contrastive studies. We take the corpus of the lexicographical project COMBIDIGILEX as example to show, how difficult it is to build a suitable data basis to study and compare linguistic phenomena in German, Spanish and Portuguese. Despite the availability of big reference corpora for the three languages (at least for written language), it is not able to obtain a comparable data basis from, because the mentioned corpora are created according to different requirements and they are also powered by disparate information systems and analyse tools. To break the status quo, we plead for increasing research infrastructures by means of compatible language technology and sharing data.

Keywords Corpus linguistics; comparative corpora; contrastive multilingual linguistics; language technologies

1. Einleitung

Korpusbasierte Analysemethoden stellen für alle sprachlichen Beschreibungsebenen interessante empirische Daten sowohl für einzelsprachige Analysen als auch für den multilingualen Sprachvergleich bereit (Hanks 2012). Korpusevidenz durch quantitative Daten in Verbindung mit entsprechenden Forschungsfragen und Hypothesen kann den Ausgangspunkt sowohl für kontrastiv angelegte Beschreibungen von Konvergenz und Divergenz als auch für anwendungsorientierte Studien für den L2-Erwerb bilden.

In den letzten zwei Jahrzehnten ist die Zahl der verfügbaren mehrsprachigen Korpora erheblich gestiegen. Sowohl Übersetzungskorpora (= Parallelkorpora) als auch vergleichbare Korpora ermöglichen empirisch angelegte kontrastive Studien mit unterschiedlichen Ansätzen und Perspektiven (Johansson 2007, S. 5 f.; Aijmer/Altberg (Hg.) 2013, S. 1 ff.; Szudarski 2018, S. 14; Trawiński/Kupietz 2021, S. 213 ff.; Meliss i. Dr.).

Im Hinblick auf die Entwicklung mehrsprachiger vergleichbarer Korpora sind internationale Initiativen wie die Entwicklung des „International Comparable Corpus“ (ICC), das derzeit 12 Sprachen umfasst (Čermáková et al. 2021), und die Initiative zur Erstellung des „European Reference Corpus“ (EuReCo) (Kupietz et al. 2020; Diewald et al. 2021; Trawiński/Kupietz 2021) zu nennen. Der Einsatz und die Entwicklung spezifischer Analyse- und Suchinstrumente ermöglichen außerdem die Durchführung groß angelegter, mehrsprachiger kontrastiver Studien auf der Grundlage vergleichbarer empirischer Daten.

Während die deutsche Sprache in vielen der genannten Initiativen vertreten ist, gibt es bislang jedoch keine institutionellen Bestrebungen, die die Einbeziehung des Spanischen und/oder des Portugiesischen in eine der oben genannten transnationalen Projekte zur Erstellung vergleichbarer mehrsprachiger Korpora vorsehen. Ausgehend von dieser Situation ist die Durchführung von Studien mit Spanisch und Portugiesisch im Kontrast zu anderen Sprachen auf einer breiten empirischen Basis nach wie vor äußerst komplex. Die Verfügbar-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

keit einer vergleichbaren empirischen Basis ist jedoch eine der unabdingbaren Voraussetzungen für sowohl inter- als auch intralinguale Studien. Um kontrastiv angelegte empirische Studien mit dem Spanischen und/oder Portugiesischen durchzuführen, ist es daher momentan nach wie vor notwendig, *ad hoc* eine vergleichbare empirische Basis herzustellen. Dabei ist mit Johansson (2007, S. 302) zu beachten, dass die geeignete Auswahl der Sprachkorpora als empirische Grundlage unter anderem von Faktoren abhängt, die mit dem Gegenstand und dem Ziel der jeweiligen Forschungsstudie und den Forschungsfragen zusammenhängen.

Das Projekt COMBIDIGILEX[1], welches den Forschungshintergrund dieses Beitrages bildet, verfolgt u. a. das Ziel, eine geeignete Methodik für die Erstellung von korpusbasierten Studien im multilingualen Kontext (z. Z. Deutsch, Spanisch, Portugiesisch) zu entwickeln, die es ermöglicht, Forschungsfragen bezüglich konvergenter und divergenter Informationen zu dem verbalen Kombinationspotenzial im Sprachkontrast durch feingranulare Untersuchungen herauszuarbeiten. Entsprechende Pilotstudien zeigen die Möglichkeiten und Grenzen der entwickelten Methodik auf (Meliss et al. (Hg.) in Vorb.) und bilden außerdem die Datengrundlage für die Entwicklung des digitalen, multilingualen, lexiko-grammatischen Informationssystems CombiDigiLex (Fernández Méndez/Mas Álvarez/Meliss 2022). Die theoretischen und methodologischen Grundlagen des Projekts verbinden korpusbasierte Analyseansätze zum verbalen Kombinationspotenzial an der Semantik-Syntax-Schnittstelle mit semantischen Ansätzen zur Bedeutungsähnlichkeit bei Verben sowie mit der kontrastiven Linguistik im deutsch-iberoromanischen Bereich, der Korpuslinguistik und der modernen Internet-Lexikographie.

Ziel dieses Beitrags ist es, zum einen die Methoden vorzustellen, die bei der Erstellung der vergleichbaren korpusbasierten Datengrundlage für das erwähnte Projekt angewendet wurden, und zum anderen die zahlreichen Herausforderungen zu diskutieren, die hierbei bewältigt werden mussten (vgl. Abschn. 2). In dem abschließenden Abschnitt 3 werden Desiderata aufgezeigt, die für zukünftige korpusbasierte Studien im multilingualen Kontext mit den besagten Sprachen neue Wege aufweisen sollen.

2. Herausforderungen

Zunächst stellt sich die Frage, wie vergleichbar Korpora unterschiedlicher Sprachen sein können und wie ein hohes Maß an Vergleichbarkeit erzielt werden kann. Das multilinguale Arbeitskorpus des COMBIDIGILEX-Projekts setzt sich zusammen aus nach unterschiedlichen Filtern zusammengestellten Subkorpora großer einzelsprachiger (Referenz-)Korpora. Folgende vier Kriterien wurden dafür verfolgt (González Ribao/Meliss/Proost in Vorb.):

1) **Medialität**: Die Auswahl der Korpusdaten beschränkt sich auf die medial geschriebene Sprache.

2) **Verteilung und Zusammensetzung** der im Korpus vertretenen Text-sorten: Das Korpus besteht aus den folgenden vier schriftsprachlichen Textsorten: Presse (P), Belletristik (BE), Wissenschaft (WI) und Gebrauchsliteratur (GL). Auf diese Weise kann der Einfluss

[1] Förderung: MINECO & FEDER (FFI2015-64476-P); vgl. https://combidigilex.wixsite.com/deutsch (letzter Zugang: 15-05-2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

der jeweiligen Textsorte auf das Kombinationspotenzial der analysierten Verben untersucht und den diesbezüglich formulierten Forschungsfragen nachgegangen werden.

3) **Zeitraum**: Die chronologische Einschränkung und Abgrenzung auf die Zeitspanne 1990–2015 hat das Ziel, eine überschaubare Menge von aktuellen Daten[2] bereitzustellen.

4) **Sprachvarietät**: Eine geografisch-politische Eingrenzung auf die europäischen Sprachvarietäten des Spanischen und Portugiesischen und die areal definierte deutsche Sprachvarietät von Deutschland hat das Ziel, das Arbeitskorpus relativ klein zu halten.

Die damit verbundene Eingrenzung bzw. Abgrenzung der großen einzelsprachlichen (Referenz-)Korpora führt zu der Erstellung entsprechender einzelsprachlicher Subkorpora (vgl. Abb. 1). Die Datengrundlage für das Deutsche setzt sich aus Texten unterschiedlicher Korpora zusammen. Hiermit wurde hauptsächlich ein hoher Grad an inhaltlicher und typologischer Übereinstimmung der deutschen Textsammlung mit dem Textangebot der entsprechenden spanischen Referenzkorpora angestrebt. Mithilfe des Korpusrecherche-, Verwaltungs- und Analysesystems COSMAS II wurde aus dem Deutschen Referenzkorpus (= DeReKo, Release 2017)[3] für das Presse-Subkorpus ein virtuelles *Ad-hoc*-Korpus erstellt, das aus ausgewählten regionalen und überregionalen Zeitungen und verschiedenen Zeitschriften besteht. Es wurde berücksichtigt, dass die Auswahl an Zeitungen und Zeitschriften für das *Ad-hoc*-Korpus inhaltlich und thematisch dem Presseteil in CREA und CORPES ähnelt[4], um die Vergleichbarkeit der Materialien für beide Sprachen weitestgehend zu garantieren. Die anderen drei textsortenspezifischen Subkorpora wurden aus den Kernkorpora (KK) des 20. und des 21. Jahrhunderts des Digitalen Wörterbuchs der deutschen Sprache (= DWDS) zusammengesetzt (Geyken 2007). Die verwendeten spanischen Korpora und die DWDS-Kernkorpora für diese drei Textsorten weisen hinsichtlich der oben erwähnten Kriterien ein hohes Maß an Vergleichbarkeit auf (González Ribao 2021, S. 62 ff.). Für das Arbeitskorpus des Spanischen wurden die zwei Referenzkorpora der königlich spanischen Sprachakademie (= RAE) herangezogen und über die integrierte Suchmaschine abgefragt. CREA wurde für den Zeitabschnitt 1990–2000 und CORPES XXI für 2001–2015 genutzt. Zur Erstellung des Arbeitskorpus für das Portugiesische wurde das Referenzkorpus des zeitgenössischen Portugiesischen (= CRPC) verwendet, das über die Rechercheplattform CQPweb abgefragt werden kann (Mendes et al. 2012).

[2] Bei Projektbeginn lagen keine aktuelleren Daten vor.

[3] Vgl. Kupietz et al. (2018).

[4] Das DWDS bietet zwar auch ein Pressekorpus an, aber für die hier besagten Studien wurde aus folgenden Gründen mit DeReKo gearbeitet. Zum einen stellt das Letztere eine größere Vielfalt an Pressetexten (auch Zeitschriften) zur Verfügung. Zum anderen kann es über COSMAS II verwaltet werden, was ermöglicht, aus dem Angebot von DeReKo ein virtuelles *Ad-hoc*-Korpus zusammenzusetzen, das dem spanischen Angebot näherkommt.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

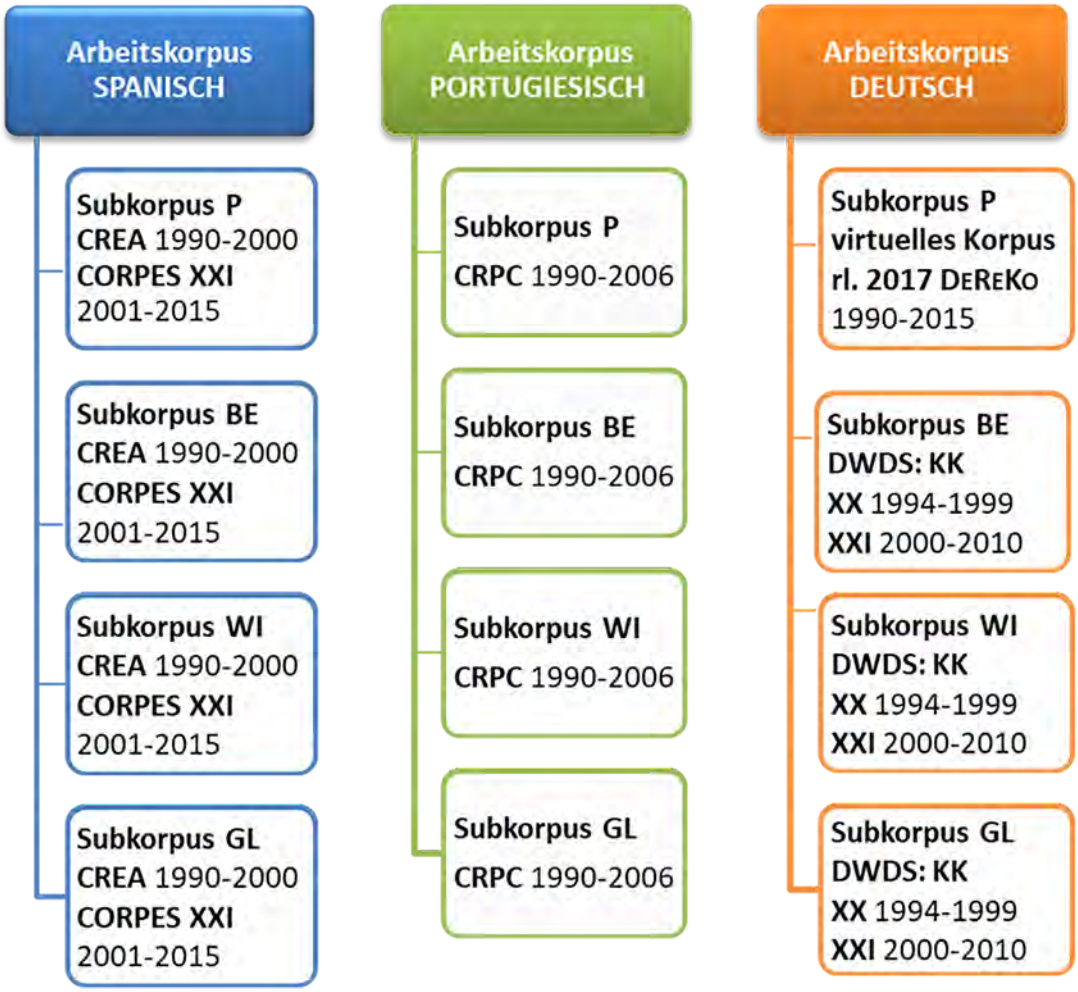


Abb. 1: Multilinguales Arbeitskorpus COMBIDIGILEX: Zusammensetzung der Subkorpora

Durch die beschriebene Methodik sollten für die drei Sprachen vergleichbare Arbeitskorpora erstellt werden, um für die projektspezifischen Forschungsfragen von COMBIDIGILEX aussagekräftige empirische Daten im multilingualen Sprachvergleich liefern zu können. Die Vergleichbarkeit der einzelnen Subkorpora konnte jedoch nur mit Einschränkung erzielt werden, weil die einzelnen Annotations-, Such- und Analysetools, über die mit den Korpora gearbeitet werden kann, nicht immer identische Funktionalitäten aufweisen. Folgende Diskrepanzen konnten aufgedeckt werden:

- **Chronologie**: Eine identische chronologische Zeitspanne konnte nicht für alle Textsorten gleichermaßen erzielt werden. Während das portugiesische Referenzkorpus auch aktuell nur Belege bis 2006 anbietet und das DWDS-Kernkorpus nur Texte bis 2010 umfasst, konnten hingegen mit dem deutschen Referenzkorpus DeReKo und dem spanischen Referenzkorpus CORPES XXI für das jeweilige Pressetextkorpus Belege bis 2015 aufgenommen werden.
- **Medialität**: Die Filterung von medial schriftlichen *vs.* medial mündlichen Texten konnte bei allen Korpora, die sowohl schriftliche als auch mündliche Daten anbieten, realisiert werden.
- **Textsorten**: Die Erstellung von textsortenspezifischen Subkorpora musste für das Deutsche durch die Kombination aus unterschiedlichen Korpora erfolgen. Da das DWDS-Zei-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

tungskorpus im Gegensatz zu den spanischen und portugiesischen Referenzkorpora weniger Variation aufweist, wurde spezifisch für das deutsche Subkorpus der Pressetexte auf DeReKo zurückgegriffen. Die metadatenorientierten, textsortenspezifischen Filterfunktionen konnten in den jeweiligen Korpora allerdings zufriedenstellend angewandt werden.

- **Varietäten**: Durch eine arealorientierte Filterfunktionalität konnten in den Referenzkorpora des Spanischen und Portugiesischen die europäischen Varianten direkt gefiltert werden. Eine Beschränkung der deutschsprachigen Korpora auf den politisch-geographischen Sprachraum Deutschland konnte in DeReKo jedoch nur durch die komplexe benutzervordefinierte Auswahl der einzelnen Textkorpora erfolgen, die gleich zu Beginn der Korpusrecherche getätigt werden muss. Für die Subkorpora aus DWDS konnte keine explizite Filterfunktion bezüglich geographischer Variation genutzt werden.
- **Tools**: Die unterschiedlichen Korpusanalysetools erlauben in den meisten Fällen keinen adäquaten Export der Ergebnisse, um auf diesen eine weiterführende qualitative Analyse anzuschließen.[5] Auch die entsprechende Visualisierung der Daten im Vergleich, dem ein hoher Nutzen für die Erkenntnisgewinnung zugeschrieben werden kann, ist oft nur schwerlich zugänglich.

Die Grundlage für die **qualitativen** und **quantitativen** Analysen bildet eine entsprechende Belegsammlung, die sich auf zufallsgenerierte Stichproben von idealerweise 100 auswertbaren Belegen pro Textsorte und lexikalischer Einheit der jeweiligen Subkorpora beschränkt. Das heißt, dass nach einer ersten Bereinigung[6] angestrebt wurde, insgesamt 400 Belege pro Lexem händisch nach vorher erstellten Kodierparametern zu analysieren.[7] Die gesamte Größe der oben erwähnten Arbeitskorpora ist daher dynamisch, denn diese wachsen mit der Anzahl der Lexeme und den entsprechenden Belegsammlungen, die zur Analyse aufgenommen werden.

Für die Analysen im multilingualen Sprachvergleich sind außerdem folgende Problembereiche zu nennen:

- **Statistik**: Die Anwendung von statistischen Methoden und entsprechende Berechnungen erweisen sich bei der Arbeit mit unterschiedlich großen Korpora oft als sehr komplex (Szudarski 2018, S. 26 f.). Hinzu kommen Pro-bleme zur Beschaffung von quantitativen Daten bei der Erstellung von Subkorpora und stratifizierten Stichproben. Bei vergleichenden Studien auf der Datengrundlage von sehr unterschiedlich großen (Teil-)Korpora ist es zudem notwendig, verschiedene Vergleichsmaße zur Berechnung anzusetzen.
- **Manuelle Analysen**: Die immer noch sehr aufwändigen manuellen einzelsprachlichen und mehrsprachigen vergleichenden Analysearbeiten erweisen sich oft als Sisyphusarbeit. Korpusbasierte und statistische Methoden erleichtern zwar unbestreitbar die Arbeit

[5] Als besonders problematisch hat sich im Fall der spanischen Korpora die Zufallsgenerierung von Samples und dessen Export sowie die quantitativen Informationen bezüglich des gesamten Korpusumfanges erwiesen. Außerdem erlaubt das entsprechende Verwaltungssystem keine Sortierung der Treffer nach dem Zufallsprinzip.

[6] Durch eine manuelle Bereinigung wurden bestimmte Belege als ungültig kodiert. Dazu wurden Merkmale wie u. a. Unvollständigkeit herangezogen. Für einige Lexeme konnten nicht immer 100 gültige Belege pro Textsorte registriert werden.

[7] Die Kodierparameter werden in González Ribao/Meliss/Proost (in Vorb.) ausführlich vorgestellt.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

durch Vorstrukturierung von Massendaten und das Erkennen von bestimmten Gebrauchsphänomenen, die linguistische Interpretation bleibt jedoch nach wie vor in den Händen der LinguistInnen (Ďurčo 2010, S. 120).

3. Desiderata

Aus den aufgezeigten Problemfeldern wird deutlich, dass es unabdingbar ist, sowohl größere Mengen variationsreicher Sprachdaten für die Erstellung von multilingualen Korpora unterschiedlichster Ausprägungen bereitzustellen als auch für korpusbasierte linguistische Studien im multilingualen Kontext in Zukunft noch mehr, bessere und benutzerfreundlichere digitale Korpustechnologien zu entwickeln und einzusetzen. Dies erleichtert nicht nur die Arbeit, sondern erhöht auch die Qualität der Ergebnisse und die Anzahl der korpusbasierten Analysen an sich.

Für multilinguale Korpusstudien wären u. a. die Entwicklung benutzerfreundlicher korpusunabhängiger Such- und Analysesoftware wünschenswert, mit der (Teil-)Korpora unterschiedlicher Sprachen und verschiedener medialer Formen gleichermaßen über eine einzige Benutzeroberfläche kostenfrei abgefragt werden können. Diese multifunktionalen Werkzeuge bzw. die Integration von verschiedenen Werkzeugen müsste neben entsprechenden Filterfunktionen zu Metadaten (einzelsprachlich und im multilingualen Kontrast) und weiteren klassischen Funktionen (Konkordanzen, Kollokationen etc.) auch u. a. folgende Funktionalitäten für alle integrierten Korpora vereinen:

a) einzelsprachliche und mehrsprachige Abfragen von Kookkurrenzen (n-Gramme etc.)

b) einzelsprachlich und mehrsprachige Abfrage von annotierten Korpusdaten (POS, Formen, Semantik, Syntax etc.)

c) unterschiedliche Strukturierungsmöglichkeiten der Daten (auch Möglichkeit der Zufallsgenerierung)

d) benutzerfreundliche Exportfunktionen der Daten

e) Angebot von unterschiedlichen statistischen Methoden zur Berechnung von Häufigkeiten nach verschiedenen statistischen Parametern

f) Möglichkeiten zur Visualisierung der Daten im Vergleich

Schritte in diese aufgezeigten Richtungen werden in unterschiedlichen Projekten und an unterschiedlichen Institutionen schon seit geraumer Zeit unternommen. Ein bekanntes Beispiel für eine solche fortgeschrittene Software ist Sketch Engine mit zahlreichen Funktionalitäten für Korpora vieler Sprachen und unterschiedlicher Größen (Kilgariff et al. 2014). Es steht auch zunehmend kostenfrei verfügbare Software, wie z. B. AntConc (Anthony 2022), für spezifische Forschungsfragen im multilingualen Kontext zur Verfügung.

Bezüglich der Entwicklung von modernen multifunktionalen Recherchesystemen soll an dieser Stelle außerdem auf KorAP verwiesen werden, welches nicht nur für das Deutsche Referenzkorpus genutzt wird, sondern auch für EuReCo (Kupietz et al. 2020; Diewald et al. 2021). In diesem Rahmen werden auch weitere benutzerfreundliche Tools entwickelt (Kupietz/Diewald/Margaretha 2020).

Dennoch besteht aktuell ein klarer Bedarf an weiteren mehrsprachigen Korpora unterschiedlichster Ausprägungen, die ein hohes Maß an Vergleichbarkeit gewährleisten (Trawiński/Kupietz 2021, S. 218). Außerdem plädieren wir u. a. dafür, für multilingual-korpus-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

basierte Studien mittels einer sprachübergreifenden Korpussuch- und Analyseplattform zielgerichtet mehr Sprach- und Korpustechnologie einzusetzen. Bedingung dafür ist u. a. der freie Zugriff auf die entsprechenden Korpusdaten. Konkret für den deutsch-iberoromanischen Sprachvergleich auf der Grundlage von großen Referenzkorpora sollten die genannten Desiderata unbedingt an die oben erwähnten schon existierenden europäischen Initiativen anknüpfen, da diese für kontrastive Studien eine bessere Ausgangslage zu versprechen scheinen.[8]

Durch die Verbindung von digitaler Forschungsinfrastruktur und humanen Ressourcen auf europäischer Ebene sollten somit auch in dem Bereich der multilingualen Korpuslinguistik Synergien verstärkt gefördert und erschaffen werden. Neben einer hohen Arbeitserleichterung für korpusbasierte sprachtheoretische Fragestellungen könnte v. a. die moderne Internetlexikographie von diesen Vorschlägen sowohl bei dem lexikographischen Prozess als auch bei der Einbindung der Daten in entsprechende Ressourcen für die unterschiedlichsten Zielgruppen und Benutzersituationen profitieren (Gouws 2021, S. 16).

Literatur

Aijmer, K./Altenberg, B. (Hg.) (2013): Advances in corpus-based contrastive linguistics. Studies in honour of Stig Johansson. Amsterdam/Philadelphia.

AntConc: freeware corpus analysis toolkit for concordancing and text analysis. https://www.laurenceanthony.net/software/antconc/ (Stand: 15.5.2022).

Anthony, L. (2022): What can corpus software do? In: O'Keeffe, A./McCarthy, M. (Hg.): The Routledge handbook of corpus linguistics. Abingdon/New York, Chapter 9.

Čermáková, A./Jantunen, J./Jauhiainen, T./Kirk, J./Křen, M./Kupietz, M./Uí Dhonnchadha, E. (2021): The International Comparable Corpus: challenges in building multilingual spoken and written comparable corpora. In: Research in Corpus Linguistics 9 (1). Special issue "Challenges of combining structured and unstructured data in corpus development", S. 89–103.

CORPES XXI: Corpus del Español del Siglo XXI. https://www.rae.es/banco-de-datos/corpes-xxi (Stand: 15.5.2022).

COSMAS II: Corpus Search, Management and Analysis System. https://www2.ids-mannheim.de/cosmas2/uebersicht.html und https://cosmas2.ids-mannheim.de/cosmas2-web/ (Stand: 15.5.2022).

CREA: Corpus de Referencia del Español Actual. https://corpus.rae.es/creanet.html (Stand: 15.5.2022).

CRPC: Corpus de Referência do Português Contemporâneo, Lisboa: Centro de Linguística da Universidade de Lisboa.
https://clul.ulisboa.pt/projeto/crpc-corpus-de-referencia-do-portugues-contemporaneo (Stand: 15.5.2022).

COMBIDIGILEX: Projekt: Kombinatorik in lexikalisch-semantischen Paradigmen im Kontrast. Empirische Studien und Digitalisierung für den Fremdsprachenerwerb in germanisch-iberoromanischen Kontexten. https://combidigilex.wixsite.com/deutsch (Stand: 15.5.2022).

[8] An dieser Stelle ist zu bedauern, dass sich die königlich spanische Sprachakademie bis jetzt weder an die Initiativen der EFNIL (European Federation of National Institutions for Language) noch an die EuReCo-Initiative angeschlossen hat. Kontrastive, korpusbasierte Studien mit dem Spanischen sind daher nach wie vor mit großen Herausforderungen verbunden, denen sich die Sprachforschenden mit unterschiedlichen Methoden und Strategien stellen müssen.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

CombiDigiLex: Digitales, multilinguales, lexiko-grammatisches Informationssystem. Prototype V.1.0.8., Santiago de Compostela: Universidade de Santiago de Compostela. http://combidigilex.usc.gal/index.php# (Stand: 15.5.2022).

DeReKo: Deutsche Referenzkorpus. https://www.ids-mannheim.de/digspra/kl/projekte/korpora/ (Stand: 15.5.2022).

Diewald, N./Bodmer, F./Harders, P./Irimia, E./Kupietz, M./Margaretha, E./Stallkamp, H. (2021): KorAP und EuReCo – Recherchieren in mehrsprachigen vergleichbaren Korpora. In: Lobin, H./Witt, A./Wöllstein, A. (Hg.): Deutsch in Europa. Sprachpolitisch, grammatisch, methodisch. (= Jahrbuch des Instituts für Deutsche Sprache 2020). Berlin/Boston, S. 287–294.

Ďurčo, P. (2010): Extracting data from corpora statistically – pros and cons. In: Ďurčo, P. (Hg.): Feste Wortverbindungen und Lexikographie. (= Lexicographica. Series Maior 138) Berlin/New York, S. 43–48.

DWDS: Der deutsche Wortschatz von 1600 bis heute. https://www.dwds.de/ (Stand: 18.3.2022).

EFNIL: European Federation of National Institutions for Language. http://www.efnil.org/ (Stand: 16.5.2022).

Fernández Méndez, M./Mas Álvarez, I./Meliss, M. (2022): Herausforderungen bei der Erstellung der multilingualen, korpusbasierten lexikografischen Ressource CombiDigiLex. In: TEISEL. Tecnologías para la investigación en segundas lenguas, Universitat de Barcelona, 1/2022. DOI: https://doi.org/10.1344/teisel.v1.36590.

Geyken, A. (2007): The DWDS corpus: a reference corpus for the German language of the 20th century. In: Fellbaum, C. (Hg.): Collocations and idioms: linguistic, lexicographic, and computational aspects. London, S. 23–41.

González Ribao, V. (2021): Mediale Kommunikationsverben. Das Zusammenspiel von Verb- und Musterbedeutung im Sprachvergleich Deutsch-Spanisch. (= Konvergenz und Divergenz 12). Berlin/Boston.

González Ribao, V./Meliss, M./Proost, K. (in Vorb.): Argumentstrukturen kontrastiv: Methodologische Grundlagen für korpusbasierte quantitative und qualitative Studien. In: Meliss, M./Mas Álvarez, I./Sánchez Hernández, P./González Ribao, V. (Hg.): Argumentstrukturmuster im Sprachvergleich. Korpusbasierte Studien zu Verben ausgewählter Paradigmen. (= Konvergenz und Divergenz). Berlin/Boston.

Gouws, R. (2021): Expanding the use of corpora in the lexicographic process of online dictionaries. In: Piosik, M./Taborek, J./Woźnicka, M. (Hg.): Korpora in der Lexikographie und Phraseologie. Stand und Perspektiven. (= Lexicographica. Series Maior 160). Berlin/Boston, S. 1–19.

Hanks, P. (2012): Corpus evidence and electronic lexicography. In: Granger, S./Paquot, M. (Hg.): Electronic lexicography. Oxford, S. 57–82.

Johansson, S. (2007): Seeing through multilingual corpora. On the use of corpora in contrastive studies. (= SCL: Studies in Corpus Linguistics 26). Amsterdam/Philadelphia.

Kilgarriff, A./Baisa, V/Bušta, J./Jakubíček, M./Kovář, V./Michelfeit, J./Rychlý, P./Suchomel, V. (2014): The Sketch Engine: ten years on. In: Lexicography: Journal of ASIALEX 1 (1), S. 7–36.

KorAP: https://korap.ids-mannheim.de/ (Stand: 15.5.2022).

Kupietz, M./Diewald, N./Margaretha, E. (2020): RKorAPClient: An R Package for accessing the German Reference Corpus DeReKo via KorAP. In: Calzolari, N./Béchet, F./Blache, P./Choukri, K./Cieri, C./Declerck, T./Goggi, S./Isahara, H./Maegaard, B./Mariani, J./Mazo, H./Moreno, A./Odijk, J./Piperidis, S. (Hg.): Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC), May 11–16, 2020, Palais du Pharo, Marseille. Marseille, S. 7016–7021.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Kupietz, M./Diewald, N./Trawiński, B./Cosma, R./Cristea, D./Tufiş, D./Váradi, T./Wöllstein, A. (2020): Recent developments in the European Reference Corpus EuReCo. In: Granger, S./Lefer, M. (Hg.): Translating and comparing languages: corpus-based insights. (= Corpora and Language in Use, Proceedings 6). Louvain-la-Neuve, S. 257–273.

Kupietz, M./Lüngen, H./Kamocki, P./Witt, A. (2018): The German Reference Corpus DeReKo: new developments – new opportunities. In: Calzolari, N./Choukri, K./Cieri, C./Declerck, T./Goggi, S./Hasida, K./Isahara, H./Maegaard, B./Mariani, J./Mazo, H./Moreno, A./Odijk, J./Piperidis, S./Tokunaga, T. (Hg.): Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, S. 4353–4360.

Meliss, M. (i. Dr.): Multilinguale Studien mit vergleichbaren Korpora: Möglichkeiten, Grenzen und Desiderata für den deutsch-iberoromanischen Kontext. In: Kongressakten IVG Palermo 2021. Jahrbuch für internationale Germanistik. (= Publikationen der Internationalen Vereinigung für Germanistik). Frankfurt a. M. u. a.

Meliss, M./Mas Álvarez, I./Sánchez Hernández, P./González Ribao, V. (Hg.) (in Vorbr.): Argumentstrukturmuster im Sprachvergleich. Korpusbasierte Studien zu Verben ausgewählter Paradigmen. (= Konvergenz und Divergenz). Berlin/Boston.

Mendes, A./Généreux, M./Hendrickx, I./Pereira, L./Bacelar do Nascimento, M. F./Antunes, S. (2012): CQPWeb: Uma nova plataforma de pesquisa para o CRPC. In: Costa, A./Flores, C./Alexandre, N. (Hg.): XXVII Encontro Nacional da Associação Portuguesa de Linguística. Textos Seleccionados 2011. Lissabon, S. 466–477.

RAE = REAL ACADEMIA ESPAÑOLA http://www.rae.es (letzter Zugang: 15-05-2022).

Szudarski, P. (2018): Corpus linguistics for vocabulary. A guide for research. London/New York.

Trawiński, B./Kupietz, M. (2021): Von monolingualen Korpora über Parallel- und Vergleichskorpora zum Europäischen Referenzkorpus EuReCo. In: Lobin, H./Witt, A./Wöllstein, A. (Hg.): Deutsch in Europa. Sprachpolitisch, grammatisch, methodisch. (= Jahrbuch des Instituts für Deutsche Sprache 2020). Berlin/Boston, S. 209–234.

Kontaktinformationen

Meike Meliss
Universidad de Santiago de Compostela
meike.meliss@usc.es

Vanessa González Ribao
Postdoc-Stipendiatin der Fritz-Thyssen-Stiftung
vanessina_gr@hotmail.com

Danksagung

Wir danken für den finanziellen Teilsupport über die Forschungsgruppe Humboldt GI-1920 (USC). An dieser Stelle bedanken wir uns auch herzlich bei den zwei anonymen Gutachtenden für ihre hilfreichen Kommentare sowie bei dem COMBIDIGILEX-Team für die Anwendung der entwickelten korpusbasierten Methodik bei den multilingualen Analysen und die entsprechenden kritischen Rückmeldungen.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Irene Renau/Rogelio Nazar

TOWARDS A MULTILINGUAL DICTIONARY OF DISCOURSE MARKERS

Automatic extraction of units from parallel corpus

Abstract This paper presents a multilingual dictionary project of discourse markers. During its first stage, consisting of collecting the list of headwords, we used a parallel corpus to automatically extract units from texts written in Spanish, Catalan, English, French and German. We also applied a method to create a taxonomy structure for automatically organising the markers in clusters. As a result, we obtain an extensive, corpus-driven list of headwords. We present a prototype of the microstructure of the dictionary in the form of a standard XML database and describe the procedure to automatically fill in most of its fields (e. g., the type of DM, the equivalents in other languages, etc.), before human intervention.

Keywords Computational lexicography; corpus-driven lexicography; discourse markers; multilingual lexicography

1. Introduction

In this paper we present *Dismark*, an ongoing multilingual dictionary project on discourse markers (DMs), especially oriented towards those that are used on written texts. We focus on the first stage of the project: the automatic extraction of the list of headwords of the dictionary (also called macrostructure, Hartmann 2001, p. 64). We also deal with the first tasks concerning the microstructure, that is, the organisation of the information in the entries and the way the different elements are connected to each other (ibid., pp. 64 f.).

We use a parallel corpus to detect DMs with similar functions in different languages (so far, in Spanish, Catalan, English, French and German), to obtain an extensive corpus-driven lemma list. This is a very different approach from traditional DMs' dictionaries, which are manually crafted based on previous dictionaries or classifications. For the manual creation of a prototype we used Lexonomy (Měchura 2017), an online dictionary software that provides functions to create, import and export database contents in the XML standard.

This project is motivated by the fact that online DM dictionaries are scarce, they tend to be outdated, incomplete, and often lack multilingual support. There are also general dictionaries that contain DMs among their entries, but they receive the same lexicographic treatment as regular lexical units. This is far from ideal as DMs, due to their functional nature, require specific solutions. Dealing with DMs means to consider practical aspects of written production and comprehension, such as punctuation, discursive order, register, multifuncionality, etc. In the first stages of our project, *Dismark* will contain general information about DMs covering the needs of a standard user in a literate society in which written documents are fundamental (Smith/Schryer 2013). In further stages, however, it will be possible to narrow down the type of dictionary to accommodate it to the more specific needs of particular groups of professionals or students.

In the following sections, we first provide a theoretical framework about the concept and categorisation of DMs (section 2), we then explain the method we used to extract the DMs

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

from corpus (section 3), we later provide a preliminary description of the microstructure of the dictionary (section 4) and, finally, we arrive to some conclusions and propose a program for future work (section 5).

2. Theoretical framework

In recent years, DMs have attracted considerable attention in linguistic research (e. g. Casado Velarde 1993; Fraser 1999; Martín Zorraquino/Portolés 1999; Pons 2001; Fischer 2006; Borreguero/López 2010). Early interest on the subject began to appear in the context of text grammar and discourse analysis (van Dijk 1973, 1978; Halliday/Hasan 1976; Halliday 1985). In these preliminary studies, DMs were described as particles used to facilitate the coherent interpretation of texts. In other words, instructions to connect the different propositions in a text and to organis e the argumentation. They are, for this reason, considered functional rather than lexical units, as they provide procedural instead of conceptual information. Notwithstanding this characterization, DMs do play an important role in written and oral communication. They not only connect and organise parts of discourse, but can also indicate subjectivity or attitudes, or may even be used to regulate the interaction between participants in communication (Fox Tree 2015). They are, thus, fundamental textual pieces which lay on an intermediate space between grammar and lexicon.

DMs are difficult to recognise and categorise (Cartoni/Zufferey/Meyer 2013). They can be single or multiword expressions, and they can pertain to different categories, such as conjunctions, adverbs, and prepositional phrases, among others. The most applied approach for the organisation is their functional similarity. Among the most found categories, one can find for instance additive connectives (*also*, *furthermore*); contrastive connectives (*however*, *nevertheless*); causal connectives (*consequently*, *for this reason*), and a large number of other categories and examples.

Different ways to categorise DMs have been discussed in discourse studies (Fuentes Rodríguez 1987; Fraser 1999; Martín Zorraquino/Portolés 1999; Pons 2001), but they have not yet been described in dictionaries with sufficient detail and precision, probably due to their complexity and discursive nature. Attempts to create extensive catalogues or dictionaries of DMs are comparatively less numerous. In Spanish, some prominent examples are Santos Río (2003), Briz (2008) and Holgado Lague (2017). For other languages, there are taxonomies in English (Knott 1996), German (Stede 2002), French (Roze 2012), Portuguese (Mendes et al. 2018) and Italian (Feltracco et al. 2016), among others. In addition, an important initiative has appeared in recent years, to integrate different resources in a large, manually curated, multilingual database of DMs (Stede/Scheffer/Mendes 2019).

Efforts for the elaboration of taxonomies and catalogues of these units have been made in the past mostly by qualitative means, often by introspection, and sometimes resorting to qualitative analysis of corpora. A well-known example of this traditional approach in Spanish is the taxonomy of DMs by Martín Zorraquino/Portolés (1999), which is also valid for other languages as well. The limitations of this methodology, however, are that it can only produce a limited number of examples per category. Comparatively, less bibliography exists regarding their computational treatment, particularly using quantitative and empirical methods. This is rather surprising, considering the advantages that such methods offer. For instance, they help to overcome the subjective bias of introspection and, with efficient automation, it is possible to process massive corpora, which may lead then to the retrieval of thousands of particular DMs and also to the potential discovery of patterns of use.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

In contrast to our present research, which is based on a lexicographic perspective, most publications in the field of computational linguistics dealing with DMs are concentrated in the area most closely related to discourse analysis (e.g., Stubbs 1996; Marku 1998; Moore 2003, Webber et al. 2019). This means that most researchers in this trend are less interested in extracting and organising full inventories of DMs than in analysing instances of texts to find cases of coherence relations expressed by these units. Both problems are of course related, but they are not the same, as one deals with types and the other with tokens. The relation is given by the fact that, to analyse DMs in particular texts, one needs some form of dictionary, and this results in the need to create this type of resources. For instance, there have been some categorization attempts using techniques such as clustering and machine learning (Alonso/Castellón/Padró 2002; Hutchinson 2005; Debortoli et al. 2016), although limited to certain types of units and consuming considerable external resources, such as manual annotation, which has the potential for a biased classification.

Regarding the specific use of parallel corpora for the study of DMs in computational linguistics, previous research is even more scarce. Some authors have used parallel corpora as a method to discover ambiguous DMs (Versley 2010; Zhou et al. 2012), and Robledo/Nazar (2018) used a clustering method from parallel corpora, but again limited to parenthetical markers and using a variety of external resources. In contrast to these methods, our current proposal is conceptually and computationally simple, more generalizable, and less dependent on external knowledge. The method presented here is a further development of ideas suggested earlier by Nazar (2021).

3. Methodology for the compilation of DMs using a parallel corpus

We propose a method to obtain an extensive inventory of the DMs of a given language, provided that a sufficiently large parallel corpus is available for that language and some other. We describe an algorithm to fully automatise all the process, starting from the corpus and finishing with a ready-to-use database. This database contains a hierarchical organisation of categories of DMs, populated with many examples in the languages under examination. In addition, our method is designed for a dynamic process, because once a first version of the database is created, it is then used to provide examples for the automatic categorisation of new DMs, thus further populating said database. These new DMs may come from other sources, not necessarily the same initial corpus.

The core idea of the method is to first separate the DMs from the rest of the vocabulary of the corpus, and then classify them according to a novel clustering algorithm. Classifying, in this case, means also finding out which are the categories, as they are not predefined. The categories are thus a product of the process, as much as the specific DMs populating them.

To facilitate future replication in other languages, we also avoid all forms of explicit knowledge of a particular language, even POS-taggers. The proposed method is thus purely statistical using only corpus as input. The only sense in which we use predetermined knowledge is regarding the names for the categories, which we borrow from Martín Zorraquino/Portolés (1999), but we consider these names can be applied with independence of the language.

The only input is thus a parallel corpus, and in our case, we used Tiedemann's (2012) Opus Corpus, which offers large samples of aligned sentences of a wide variety of languages and

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

genres. This material is freely available in TMX format, which specifies the alignment of translation segments (TS), a unit of measure that typically corresponds to a sentence. There are circa 30 files per language pair in the case of European languages, and each file compresses large samples of texts (circa 3,500 million tokens) of a certain genre or discipline. The corpus is representative of a great variety of written genres.

Oral speech is only indirectly represented in files containing literature and TV subtitles, which also offer large samples of general vocabulary.

The method can be synthesised as follows.

3.1 Extracting DMs from corpus

DMs are automatically separated from the rest of the vocabulary using a co-occurrence association measure that feeds an entropy model. DMs are visible because they show a particular distribution in the corpus, a characteristic pattern that is a consequence of the fact that they are independent of the content of the text in which they occur. In operational terms, this means that their occurrences show a uniform distribution, with a very wide, non-restrictive set of co-occurring words. We say they are uninformative because they cannot be used to predict the occurrence of other lexical units. In contrast, a more informative lexical unit could be *democracy*, as it shows a clear pattern of co-occurrence with a set of words such as *respect*, *freedom*, *rights*, and so on. In contrast, the word *anyway* does not have these "friends", as it only has a functional value. This difference is measured by coefficient (1) where x is a DM candidate and Rx the set of its co-occurrences.

(1) $$I(x) = \frac{\log_2 \sum_{i=1}^{n} R_{x,i}}{\log_2 |m(x)|}$$

The symbol $m(x)$ refers to the contexts candidate x, and $R(x,i)$ is the frequency of the word in position i of the ranked list of the n most frequent words that co-occurring with x in the same sentences (in our experiments, $n = 20$). In one extreme, such coefficient will produce a very low score for function words such as articles, conjunctions, prepositions, etc. At the opposite extreme of this continuum, the most specialised vocabulary units begin to appear, because these are the ones that will typically point to a limited set of other units. An arbitrary threshold k determines if x is classified as a lexical or functional unit. For illustration, consider Figures 1 and 2, showing the co-occurrence profile of the Spanish word *electricidad* ('electricity') and *de todas maneras* ('anyway'), respectively. One can see the different shapes of both curves, the first one having a greater surface under the curve. It should be noted that the method could also be of interest for specialised lexicography because it may be implemented as a term-extractor, as suggested by Nazar/Lindemann (2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

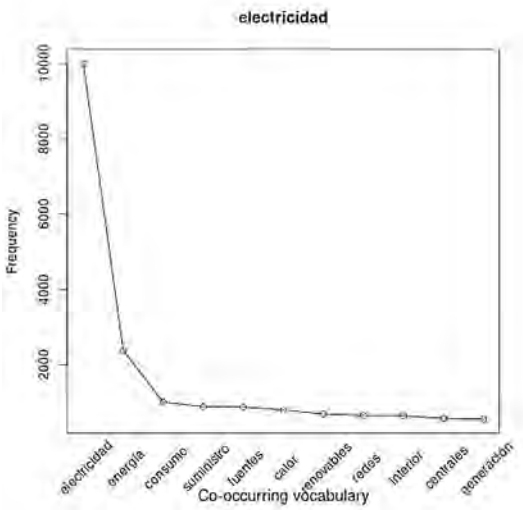


Fig. 1: Co-occurrence profile of the Spanish lexical unit *electricidad* ('electricity')

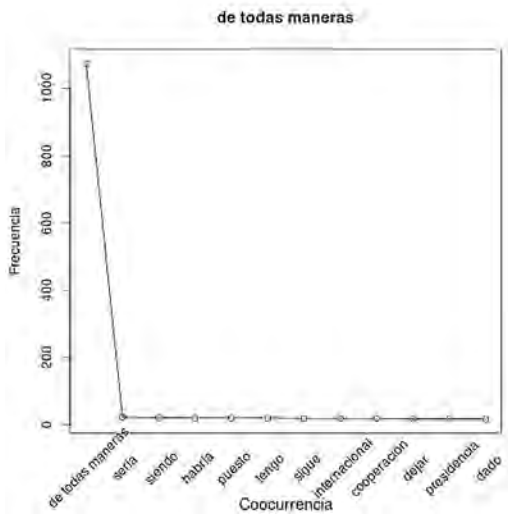


Fig. 2: Co-occurrence profile of the Spanish DM *de todas maneras* ('anyway')

3.2 Clustering DMs

We developed a clustering algorithm that uses the equivalence of the DMs in another language as a similarity measure, hence the parallel corpus. This is effectively to use the parallel corpus as semantic mirror. For instance, *nevertheless* and *however* can be considered similar because they share the same equivalences in a second language (e. g., *sin embargo* or *no obstante*, in the case of Spanish). To find the equivalences in the parallel corpus, we used an association coefficient based on the co-occurrence of DMs in the aligned sentences (2).

(2) $$A(MD_{es,i}, MD_{ca,j}) = \frac{f(MD_{es,i}, MD_{ca,j})}{\sqrt{f(MD_{es,i})} \cdot \sqrt{f(MD_{ca,j})}}$$

Once with the list of aligned DMs at hand, the clustering algorithm proceeds as follows: it takes the pairs of aligned DMs one by one, e. g. *por esa razón* and *for that reason*. If in a subsequent pair the English DM is repeated, as in the case of *por esta razón* ~ *for that reason*, then it is assumed that *por esa razón* and *por esta razón* are equivalent, that is, they have the same function and can be used in the same context. We see no need, at this point, to exploit lexical or orthographic similarity here but in any case, that is a possibility we leave for future work. If the DMs are similar, they form a new cluster. For illustration, consider a more advanced stage in this process, in which we have a situation such as the one depicted in Figure 3, with *por esta razón* already being a member of a previously formed cluster containing units such as *por ese motivo* or *por este motivo*. In such a case, the newly arrived DM *por esa razón* is added to said cluster. The process finishes when there are no more DM pairs to process.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

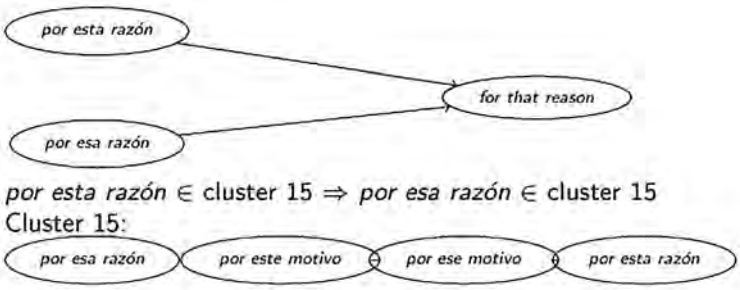


Fig. 3: A moment of the DMs clustering process

3.3 Labelling the clusters

The previous step results in several clusters of similar DMs in each language, but the system is not able at this point to produce a name for these clusters. At this point they are instead only identified with numerical codes. To give these clusters a meaningful name, we used the names of the categories in the taxonomy by Martín Zorraquino/Portolés (1999). Using the few examples they provide for their categories, we can automatically find the match with our clusters and tag them accordingly (3). Also, as all clusters are aligned by language (we keep the initial alignment obtained from the parallel corpus), the same labels are also used for the rest of the languages.

(3) $$sim(MZP_p, CMD_q) = \frac{|M\vec{Z}P_p \cap C\vec{M}D_q|}{|M\vec{Z}P_p|}$$

3.4 Populating the taxonomy with new DMs

Once a basic taxonomy of DMs is built this way, it is then used to classify new DMs in a recursive manner. The algorithm will first classify a DM candidate by language, it will then decide if it is effectively a DM and, if this is the case, it will assign a category to it. For both tasks we used the initial parallel corpus: if a Spanish candidate is a genuine DM, its condition will be signaled by the parallel corpus, because it will be associated with English DMs of the corresponding category. For instance, given a new candidate in Spanish such as *de la misma manera*, we will find that in the Spanish-English parallel corpus this appears aligned with already known English DMs such as in *the same way*, *likewise*, *similarly*, etc. We must conclude, then, that 1) the Spanish candidate is indeed a DM, and 2) it belongs to the same category as its English counterparts. Here lies also the possibility of discovering polyfunctional cases, i. e., the possibility that this Spanish DM is also associated with a different group of English DMs but, again, we leave that challenge for future research.

3.5 Evaluation of the taxonomy of DMs

As a result of this method, now we have obtained 619 candidates for Spanish, 733 for English, 556 for French, 677 for German, and 312 for Catalan, all distributed in 70 different functional categories. The taxonomy of DMs can be consulted at http://www.tecling.com/dismark (last access: 26-05-2022). A campaign for the manual evaluation of the whole col-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

lection was undertaken with the collaboration of a group of linguists that are native speakers of each of the languages, with two or three linguists per language. The revision involved periodic discussions between members of the different teams, to keep a uniform criterion in all languages.

The evaluation was conducted in two phases. The first one was to determine precision, defined as the proportion of correct DMs found in the newly created DM taxonomy. The second phase, in turn, was to estimate recall, defined as the proportion of DMs that exist in a language that are included in said taxonomy.

For the evaluation of precision, we reviewed all DMs contained in the taxonomy counting the number of cases in which a) an element is not genuine DMs; b) a multi-word DM was not correctly segmented (typically missing an initial or final part) or c) the element is actually a DM but it appears in the wrong cluster or category. The revision revealed that the percentage of errors is less than 5% in all languages except in German, where we found 16% false positives, mostly with segmentation faults. In terms of precision, we believe this result is sufficiently accurate to constitute the core for a list of headwords of the dictionary.

For the evaluation of recall, the method we devised was to obtain random samples of texts and find the proportion of DMs that are in those texts and not in the DM taxonomy, divided by the sum of said number and the total count of DMs in those texts that do also appear in the taxonomy. In a sample of ten texts per language, 88% of the DMs were already documented in our database. This does not translate directly into a measure of recall, but it indicates that at least we have the majority of the most frequent exemplars.

4. Preliminary lexicographic proposal

As stated in section 1, a first stage of the *Dismark* project contemplates to create a core of DMs units and microstructural information. The target users of the dictionary are, in this stage, professional communicators such as journalists, screenwriters, translators, lawyers, scientists, etc. (Schriver 2012) and college students in need of acquiring expertise in communication as part of their professional formation (Lea/Street 2006), e. g., students of Journalism, Law, Translation, etc. All these users share common needs, for example, what a specific DM is used for, how should they use punctuation, the orthography of the DM, etc. A second phase of the project contemplates the creation of sub-products, such as a specific version of the dictionary designed for lawyers or journalists.

The dictionary is unidirectional (Atkins/Rundell 2008, p. 40), with Spanish as the target language, and equivalents in English, French, German, and Catalan. The microstructure has the following types of information (see some sample entries in the online prototype: https://www.lexonomy.eu/#/dismark, last access: 05-26-2022):

- **Headword.** As the dictionary is made from scratch for the Internet, the lemmatisation does not contain any change of order, typical from the constraints of the alphabetic order in paper dictionaries. Thus, *aun así* 'still' is lemmatised *aun así* and not *así, aun*.
- **Type of DM.** The different types of DM are categorised according to Martín Zorraquino/Portolés (1999). This field will have a hyperlink to an external webpage containing extended information about the type of DM.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

- **Register.** We added this field to separate standard from formal DMs. As the dictionary is focused on functional writing, there are not many cases of DMs used in colloquial language.
- **Function.** In this section, we synthetically describe the function of the DM. An extended explanation of the function of the DM is already offered as hyperlink in the *Type of DM*. In this field, we want to cover the need of the user to obtain a quick and clear explanation.
- **Examples.** We provide 1–2 examples of usage, containing at least two sentences, in order to provide enough discursive context. We also provide the source of the examples, which can be different corpora or obtained from documentation or the Internet.
- **Punctuation and position.** We provide the patterns of punctuation and position that the user can find when using or reading the DM. Patterns are expressed by the punctuation sign before and/or after the DM. For example, for *sin embargo* 'however', two common patterns are:

  . *Sin embargo*,
  ; *sin embargo*,

  This allows to solve other orthographic doubts, such as capitals or blank spaces.

  Each pattern is complemented by one or more *Examples*.
- **Spanish equivalents.** A list of all DMs of the same type of the headword are offered here. These groups have been automatically extracted, as explained in the previous section, but are later manually revised. Each DM in this list contains a link to the correspondent entry.
- **Translations** to Catalan, English, French and German. The group of equivalent DMs in these languages are offered. They will be also linked to the multilingual part of the dictionary.

All types of information detailed in this list required expert human supervision. However, most of it can be automatically filled in, e. g., the list of headwords, the types of DMs, the equivalents and the patterns of punctuation and position. As for the examples, a random sample of corpus concordances of each type of pattern is added to the field, so that the lexicographer can easily select convenient examples. All this information can be automatically added, as Lexonomy allows us to work with independent XML files that can be uploaded to the database.

Figure 4 shows one of the entries of the sample prototype, *sin embargo* 'however'.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

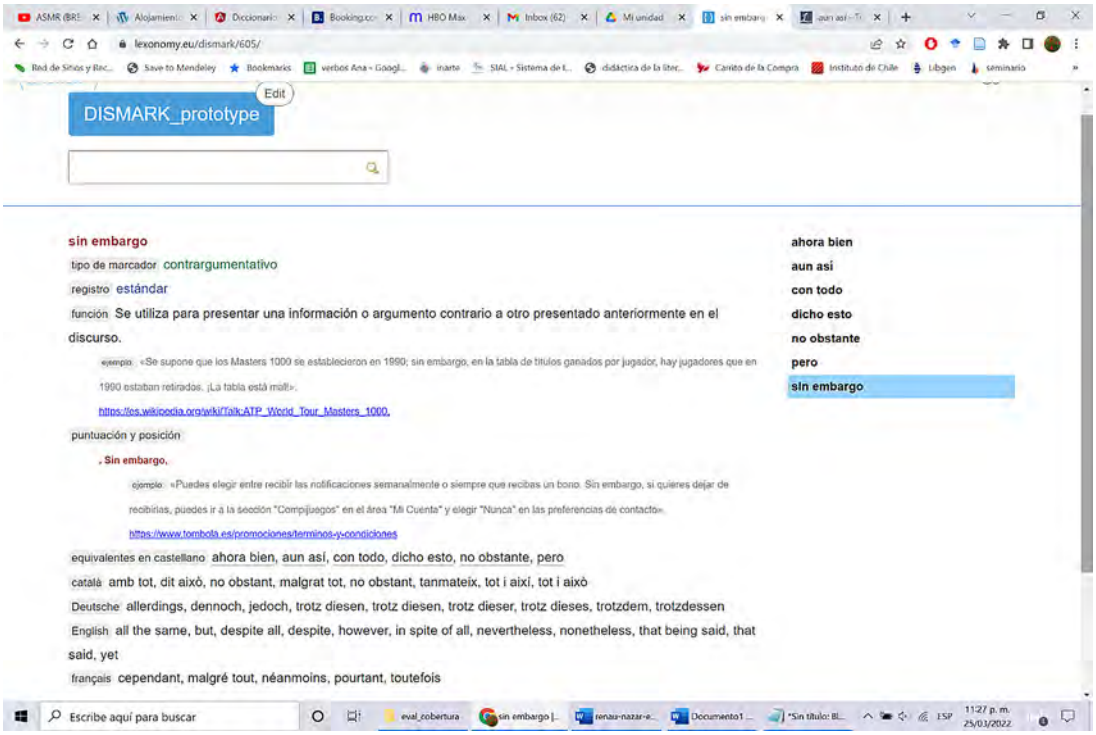


Fig. 4: Example of a *Dismark* entry in the sample prototype

5. Conclusions and further steps

In this paper, we presented our first steps towards a corpus-driven online dictionary of DMs with inter-linked entries in five languages. The method of extraction of DMs from a parallel corpus has enough precision and recall obtaining a large list of headwords for the dictionary that we are planning.

There are different tasks to be addressed in the immediate future of this project. We have to test the prototype with users and, after validation, we have to prepare a final version. There is also work to do in describing each type of DMs present in the dictionary, which will not be part of the dictionary itself, but it will be connected to it by hyperlinks. In relation to this, another important aspect to address is the design of the mediostructure (Hartmann 2001, pp. 65 f.), that is, the system of cross-references connecting different entries, parts of the dictionary with external resources, etc. We also must address the problem that some DMs can have multiple functions, as Cartoni/Zufferey/Meyer (2013) show. Finally, and as already mentioned, a long-term project will be to create sub-types of the same dictionary to address the specific needs of different types of users.

References

Alonso, L./Castellón, I./Padró, L. (2002): Lexicón computacional de marcadores del discurso. In: Procesamiento del Lenguaje Natural 29, pp. 239–246.

Atkins, S./Rundell, M. (2008): The Oxford guide to practical lexicography. Oxford.

Borreguero, M./López, A. (2010): Marcadores del discurso: De la descripción a la definición. Madrid/Frankfurt a. M.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Briz, A./Pons, S./Portolés, J. (coords.) (2008): Diccionario de partículas discursivas del español. http://www.dpde.es.

Cartoni, B./Zufferey, S./Meyer, T. (2013): Annotating the meaning of discourse connectives by looking at their translation: the translation spotting technique. In: Dialogue and Discourse 4 (2), pp. 65–86.

Casado Velarde, M. (1993): Introducción a la gramática del texto del español. Madrid.

Debortoli, S./Müller, O./Junglas, I. A./vom Brocke, J. (2016): Text mining for information systems researchers: an annotated topic modeling tutorial. In: Communications of the Association for Information Systems 39 (1), pp. 1–30.

Feltracco, A./Jezek, E./Magnini, B./Stede, M. (2016): LICO: A lexicon of Italian connectives. In: Proceedings of Third Italian Conference on Computational Linguistics (CLiC-it 2016) & Fifth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2016), Napoli, Italy, December 5–7, 2016, volume 1749 of CEUR Workshop Proceedings. CEUR-WS.org.

Fischer, K. (ed.) (2006): Approaches to discourse particles. Amsterdam.

Fox Tree, J. E. (2015): Discourse markers in writing. In: Discourse Studies 17 (1), pp. 64–82.

Fraser, B. (1999): What are discourse markers? In: Journal of Pragmatics 31, pp. 931–952.

Fuentes Rodríguez, C. (1987): Enlaces extraoracionales. Sevilla.

Halliday, M. A./Hasan, R. (1976): Cohesion in English. London.

Halliday, M. A. K. (1985): An introduction to functional grammar. London.

Hartmann, R. R. K. (2001): Teaching and researching lexicography. Harlow.

Holgado Lage, A. (2017): Diccionario de marcadores discursivos para estudiantes de español como segunda lengua. New York.

Hutchinson, B. (2005): The automatic acquisition of knowledge about discourse connectives. PhD thesis. Edinburgh.

Knott, A. (1996): A data-driven methodology for motivating a set of coherence relations. PhD thesis. Edinburgh.

Lea, M. R./Street, B. V. (2006): The "academic literacies" model: theory and applications. In: Theory Into Practice 45 (4), pp. 368–377.

Martín Zorraquino, M. A./Portolés, J. (1999): Los marcadores del discurso. In: Bosque, I./Demonte, V. (eds.): Gramática descriptiva de la lengua española, Vol. 3. Madrid, pp. 4051–4213.

Měchura, M. B. (2017): Introducing lexonomy: an open-source dictionary writing and publishing system. In: Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of the eLex 2017 Conference, 19–21 September 2017, Leiden, The Netherlands.

Mendes, A./del Rio, I./Stede, M./Dombek, F. (2018): A lexicon of discourse markers for Portuguese – LDM-PT. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). Miyazaki, Japan.

Moore, J. D./Wiemer-Hastings, P. (2003): Discourse in computational linguistics and artificial intelligence. In: Graesser, A. C./Gernsbacher, M. A./Goldman, S. R. (eds.): Handbook of discourse processes. London.

Nazar, R. (2021): Automatic induction of a multilingual taxonomy of discourse markers. In: Kosem, I. et al. (eds.): Electronic lexicography in the 21st century: post-editing lexicography. Brno, pp. 440–454.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Nazar, R./Lindemann, D. (2022): Terminology extraction using co-occurrence patterns as predictors of semantic relevance. In: Proceedings of the Workshop on Terminology in the 21st century (Term21), LREC 2022, Marseille, France.

Pons, S. (2001): Connectives/discourse markers. An overview. In: Quaderns de Filologia. Estudis Literaris 6, pp. 219–243.

Robledo, H./Nazar, R. (2018): Clasificación automatizada de marcadores discursivos. In: Procesamiento del Lenguaje Natural (61), pp. 109–116.

Roze, C./Danlos, L./Muller, P. (2012): LEXCONN: a French lexicon of discourse connectives. In: Discours – Revue de linguistique, psycholinguistique et informatique. Laboratoire LATTICE, 2012. Multidisciplinary perspectives on signalling text organisation, pp. 1–15. https://hal.inria.fr/hal-00702542.

Santos Río, L. (2003): Diccionario de partículas. Salamanca.

Schriver, K. (2012): What we know about expertise in professional communication. In: Berninger, V. W. (ed): Past, present, and future contributions of cognitive writing research to cognitive psychology. New York, pp. 275–312.

Smith, D. E./de Schryer, C. F. (2013): On documentary society. In: Bazerman, Ch. (ed.): Handbook of research on writing. Amsterdam/New York, pp. 113–117.

Stede, M. (2002): DiMLex: a lexical approach to discourse markers. In: Lenci, A./Tomaso, V. D. (eds.): Exploring the lexicon – theory and computation. Alessandria.

Stede, M./Scheffer, T./Mendes, A. (2019): Connective-Lex: a web-based multilingual lexical resource for connectives. In: Discours – A Journal of Linguistics, Psycholinguistics and Computational Linguistics 24. https://journals.openedition.org/discours/10098.

Stubbs, M. (1996): Text and corpus analysis. Oxford.

Tiedemann, J. (2012): Parallel data, tools and interfaces in OPUS. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). Istanbul, pp 2214–2218.

van Dijk, T. (1973): Text grammar and text logic. En Studies in Text Grammar. Dordrecht, pp. 17–78.

van Dijk, T. (1983): La ciencia del texto: un enfoque interdisciplinario. Barcelona.

Versley, Y. (2010): Discovery of ambiguous and unambiguous discourse connectives via annotation projection. In: Proceedings of the Workshop on Annotation and Exploitation of Parallel Corpora (AEPC). Tartu, pp. 83–92.

Webber, B./Prasad, R./Lee, A./Joshi, A. (2019): The Penn Discourse Treebank 3.0 annotation Manual. tech report. University of Pennsylvania.

Zhou, L./Gao, W./Li, B./Wei, Z./Wong, K.-F. (2012): Cross-lingual identification of ambiguous discourse connectives for resource-poor language. In: Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012). Mumbai, pp. 1409–1418.

Contact Information

Irene Renau
Pontificia Universidad Católica de Valparaíso, Chile
irene.renau@gmail.com

Rogelio Nazar
Pontificia Universidad Católica de Valparaíso, Chile
rogelio.nazar@pucv.cl

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Chris A. Smith

ARE PHONESTHEMES EVIDENCE OF A SUBLEXICAL ORGANISING LAYER IN THE STRUCTURE OF THE LEXICON?

Testing the OED analysis of two phonesthemes with a corpus study of collocational behaviour of *sw-* and *fl-* words in the OEC

Abstract Phonesthemes (Firth 1930) are sublexical constructions that have an effect on the lexico-grammatical continuum: they are recurring form-meaning associations that occur more often than by chance but not systematically (Abramova/Fernandez/Sangati 2013). Phonesthemes have been shown (Bergen 2004) to affect psycholinguistic language processing; they organise the mental lexicon. Phonesthemes appear over time to emerge as driven by language use as indexical rather than purely iconic constructions in the lexicon (Smith 2016; Bergen 2004; Flaksman 2020). Phonesthemes are acknowledged in construction morphology (Audring/Booij/Jackendoff 2017) as motivational schemas. Some phonesthemes also tend to have lexicographic acknowledgment, as shown by etymologist Liberman (2010), although this relevance and cohesion appears to be highly variable as we will show in this paper.

This paper seeks to compare two phonesthemes in a combined lexicographic and corpus study with a view to testing the results obtained. **Firstly**, following Smith (2016) which identified 11 semantic categories of *fl-* words in the OED, we analyse the OED entries for 245 *sw-* monomorphemes with a view to carrying out a key word analysis and a semantic trait analysis. The 245 monomorphemes have a total of 469 senses out of which 330 can be classified into 18 recurring semantic traits in Table 1.

semantic traits based on OED key words	**number of senses carrying the trait**
sway sweep swish	78
strike blow swipe	56
pressure swell swathe	57
sway swagger boast	11
compact cluster agitated	7
big fellow	4
flame burn waste	10
deceive sway swindle	11
faint swoon agitated	18
cool dark	7
drink	19
surface	9
hollow	10
exchange swap	6
labour toil sweat	12

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

semantic traits based on OED key words	number of senses carrying the trait
deviate deflect	6
sound	9
Total	330

Table 1: Lexicographic behaviour of sw- senses in the OED

Then, in a **second step**, the comparison between the OED analysis of *fl-* and *sw-* monomorphemes shows that *sw-* words appear less likely to undergo any semantic change and therefore appear to be less indexical. In the light of these differing lexicographic behaviours, we aim, in a third step, to analyse the collocational behaviour of some common phonesthemic verbs carrying *fl-* and *sw-*. Collocational behaviour via a collexeme analysis will enable us to identify combinatorial patterns of use. For the study, we use the very large contemporary (2 billion words) OEC corpus (2000–2005) using Sketch Engine (Kilgarriff et al. 2004). The results of the compared analysis allow us to discuss whether phonesthemes are actual (sub)lexical "chunks" deserving of a lexical status, or whether they belong to larger phraseological "chunks" or units. This question raises the issue of the architecture of the lexico-grammatical continuum, the "constructicon": does the constructicon accommodate or require a sublexical layer?

What are the repercussions for lexicography and phraseology?

Keywords Phonesthemes; analogy; collocational behaviour; OED; OEC; phraseological chunks

1. Lexicography challenges: diachronic, usage based, cognitive lexicography

1.1 The OED as a historical emergent dictionary

1.1.1 Challenges and revision: using the OED as a usage-based historical database

Thanks to its philological beginnings, the OED is well known to be the most extensive historical dictionary of English (see Considine 2016; Mugglestone 2009; Brewer 2009, 2016; Paton 1995), and is regarded as an exceptional source of diachronic information, including etymological, morphological, semantic, diasystematic, and frequency data. It is used frequently for diachronic lexical analysis, such as Allan (2012), Durkin (2016a, b).

The challenges facing the OED and its revision process are now that of a hybrid evolving database. Indeed, as Brewer (2016) explains the OED3 is constantly under revision, "blending different versions of the dictionary in a proportion that changes every quarter as the revision progresses; currently the mix is roughly one-third revised third edition and two-thirds unrevised second edition". The revision process of the electronic version of the OED is gearing towards a-based historical emergent dictionary, aiming to improve systematicity in definitions, etymology, labelling, checking of attestation dates. The revision process stages in Table 2 are clearly outlined by chief editor Michael Proffitt on the OED blog page, which aims to provide transparent information for users.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

OED Revision Project	Total Progress to date	Progress targets for 2021/22
Entries rewritten or added	149,747	3,200
Senses rewritten or added	460,846	14,000
New senses added	169,280	2,500
Quotations added	1,315,621	
Etymologies added	75,623	
Variant spellings added	264,064	

Table 2: OED revision progress update

1.1.2 Usage based lexicography and cognitive lexicography

The challenges of a usage-based lexicography coincide with those of usage-based semantics, and usage-based theories of lexical semantics. Osterman (2015), Geeraerts (2016) call for a more integrated cognitive lexicography, that takes into account the users' mental lexicon. For Rundell/Atkins (2008, p. 48) the objective of a lexicographer is to identify **norms and typicality** (see also Hanks 2013):

> If our goal is to provide 'typifications', then how do we know whether a given utterance is typical (and therefore worth describing) or merely idiosyncratic (and therefore outside our remit)? A typical linguistic feature is one that is both *frequent* and *well-dispersed*. Any usage which occurs frequently in a corpus, and is also found in a variety of text-types, can confidently be regarded as belonging to the stable 'core' of the language

For Rundell/Atkins (2008, p. 280) the goal is to accommodate a certain degree of fuzziness since the lexicon is ever-changing: Therefore, the role of lexicography is to present norms (and exploitations in Hanks 2013) for users. Typicality is an important factor:

> A prototype approach to WSD has two major advantages over the classical model: It reflects the way people create meanings when they communicate, and thus it goes with the grain of the language, and accommodates creativity and fuzziness. It makes the lexicographer's task more manageable, because it allows us to focus on the prototype and its common exploitations, rather than requiring us to predict and account for every possible instantiation of a meaning.

For the historical OED, it becomes more complex to integrate indicators of typicality outside of the current time frame. The inclusion of obsolete and infrequent words specifically, which are often disregarded in other dictionaries, does call into question the notion of typicality; norms change over time for different communities of speakers. How to improve the feedback loop between lexical semantics and lexicography? This is the hope expressed by Geeraerts (2016, p. 438) for "a constant confrontation with the facts of linguistic usage draws lexicography and lexical semantics together".

How does the structure of the usage-based lexicon influence the lexicographical information in dictionaries? These are some of the challenges of usage-based lexicography and pattern analysis. Signs of a move towards cognitive aspects of lexicography can be identified in the interest for what are coined foundation words by Mickael Proffitt in the OED blog post titled *The Oxford English Dictionary: focus areas and goals for 2021*,[1] "words

[1] https://public.oed.com/blog/the-oed-2021/ (last access: 2022-04-08).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

with the greatest longevity and frequency, and which exhibit the greatest historical, semantic, and cultural complexity". [...]. The interest lies in their "**elasticity and adaptiveness in their relationships** with other words, by forming many compounds and phrases", i. e. in the structuring role in the lexicon.

1.2 Background on phonesthemes and their function in the lexicon

1.2.1 Defining phonesthemes

Phonesthemes are frequently consonant clusters at the onset of short words in English such as *fl-* (*flush*, *flap*, *flit*, *fly*) and *sw-* (*swish*, *swap*, *swindle*, *swoop*), or *sp-*, *sn-*, etc. Phonesthemes are sublexical units that resemble affixes in that they don't exist independently from the elements they combine with. They are however distinguishable from affixes in that they are not systematic nor are they considered to be morphemic. This is the main limitation of phonesthemes in traditional morphology, if we accept the traditional view of word formation as composition into morphemes. Phonesthemes are therefore viewed as sublexical associations of meaning and form which are not systematic and call into question the traditional building block vision of the lexicon. The second limitation and difficulty pertaining to phonesthemes is the difficulty in identifying the semantic associations of the form-meaning pairing. Phonesthemes are very hard to pin down, despite many attempts to identify a core "sense" (Abramova/Fernandez 2016; Abramova/Fernandez/Sangati 2013).

1.2.2 The function of phonesthemes in the lexicon

Firth (1930, p. 184) defines phonesthemes as "[p]airings of sound-meaning that are not componential or systematic". The role of phonesthemes has been identified by Firth as early as 1930, and has been recently rediscovered via psycholinguistic research (Bergen 2004). According to Bergen (2004, p. 293) "[F]orm-meaning pairings that crucially are better attested in the lexicon of a language than would be predicted, all other things being equal". Experimental studies have shown that phonesthemes affect mental processing, that is that they have a quantifiable effect in structuring the mental lexicon. This discovery has sparked computational studies into phonesthemes, with Otis/Sagi (2008, p. 65) defining a phonestheme as: "a submorphemic unit that has a predictable effect on the meaning of a word as a whole". However, this computational perspective has come under criticism from cognitive semiotics. For Pleyer et al (2017), phonesthemes should not be viewed as a purely statistical phenomenon and are to be related to cognitive transmodality in that they are **transmodal** signs (ie).

Etymologist Anatoly Liberman (2008, 2010a, 2010b) recognizes that phonesthemes **regulate etymology**. According to Liberman (2010a, p. 251) etymological ties are affected by analogical attractions; for instance *sleazy* and *glaive* have evolved due to analogy with phonesthemes *sl-* and *gl-*: "*sleazy* may have acquired its present day meaning under the influence of *sl-*, whereas *glaive* may have come to mean 'sword' rather than 'spear' because glâ•, suggests glistening". Other instances of these shifts can be traced to words carrying the phonesthemes *fl-* such as *flatter*, *flute*, cited by Liberman 2013).[2] Liberman (2010a, p. 257)

[2] The origin of *flatter* has been hotly contested. I support the hypothesis that the word was coined in Germanic and meant "flutter around the person whose favors one wishes to obtain," with the French verb having been borrowed from Middle English. *Flutter*, *flitter*, and *flatter* begin with the

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

argues that the lexicon is affected by a multitude of such paradigmatic ties which render impossible the tracing back to a single etymological source word:

> Finally, the period of "first words" is an uninspiring construct. There have always been many words that influenced one another, people have always had neighbors from whom they borrowed words, and conflicting impulses have always been at crosspurposes. There never was a beginning. After all, we are not characters in Kipling's *Just So Stories*.

Lexicographic studies of phonesthemes have been carried out based on the Oxford English Dictionary (Smith 2016, 2019).

1.2.3 Phonesthemes in a constructional view of the lexicogrammatical continuum

An alternative framework is provided by constructionist morphology (Goldberg 2006; Audring/Booij/Jackendoff 2017) which views associations of form and meaning as constructions. Constructions can be purely motivational schemas rather than generational schemas, thus accounting for the non-systematicity of form-meaning pairings.

This recent framework is therefore attractive for the description and formal analysis of form-meaning pairings that are either phraseological or sublexical. Constructional theory allows for the existence of constructions of different sizes and complexity throughout the lexicogrammatical continuum. Such constructions represent different sizes of constructions (from single word to multiword and from single word to sublexical layer). Kwon/Round (2015, p. 2) argue it is necessary to reevaluate the status of phonesthemes and to question whether they are actually different to morphemic units: "according to what criteria, if any, do phonaesthemes distinguish themselves from non-phonaesthemic, stem-building elements?".

1.3 Purpose of this paper

From a usage-based perspective, the role of phonesthemes in organising the mental lexicon is likely to carry over into the institutionalised lexicon. According to language change theories, change is affected by the usage based combinatorial behaviour of words in discourse (Bybee 2013). The assumption is that the phonesthetic attraction (Bolinger 1965) via analogical remodelling or remotivation over time is likely to affect semantic change in the lexicon itself. A historical dictionary such as the OED provides a usable database. Corpus studies have shown there is evidence of synonym **clustering** of phonesthemic words, as well as collocational clustering.

There are two confounding issues to the question of what drives change: 1) what is the role of repetition and frequency? and 2) what is the role of qualitative salience (the relative position of a word within the field of its competitors)? **Is change driven by the overall frequency of lexical items only, or is it driven by frequency within a form-meaning paradigm (onomasiological space)?** A third question is if clustering leads to chunking of

group *fl-* that we find in *flute*. In English, contrary to German, *flute* left a jeering echo. Rather probably, *flout* has been taken over from Middle Dutch. In Modern Dutch, *fluiten* has the expected sense "whistle; play the flute," but many centuries ago it also meant "mock, jibe" (Liberman 2013) https://blog.oup.com/2013/07/flute-word-origin-etymology/.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

these forms (sublexical and phraseological)? In other words, if phonesthemes do not simply exist but develop, emerge through usage, how can we track this process? Does the lexicon require a sublexical/submorphemic layer, of which there are already signs of acknowledgment in the OED, notably in the role of assigning sources of semantic shift and analogical change in the lexicon.

2. Studying *fl-* and *sw-* words in the OED and semantic shift

2.1 Protocol for the lexicographic analysis of phonesthemes

2.1.1 Key word analysis in the OED and conceptual categories

Smith 2016 initiated the protocol for the analysis of fl- monomorphemes using the OED. The protocol has also since been applied to morphemic affixes (*-age*, *-some*) in an effort to parareplicate the experiment (Smith 2018, 2020). This protocol has now been applied to *sw-* words in an effort to test the results on further phonesthemes. The objective of the protocol is not to assume key words are directly correlated to the meaning of the phonestheme, but relies on **key words** as an indicator of lexicographic cohesion. Instead of using key words as absolute indicators, they function as relative indicators in the analysis of change. Of course, one of the main drawbacks in the inconsistency in the definitions as the OED review process continues, thereby creating fluctuation in the revised entries versus the non-revised entries as mentioned previously.

In table 3 we provide the lexicographic treatment applied to *sweep* first attested [1300] and its many senses based on the key word analysis. This strategy was applied to all *sw-* monomorphemic words in the OED.

sweep, v.	**date**	**Definition**	**key words**
sweep, v.	1300	Senses with that which is removed or moved along as the object, and derived uses. To remove, clear away, off (etc.) with a broom or brush, or in a similar way by friction upon a surface; to brush away or off.	SWAY BRUSH IMPETUS
sweep, v.	1400	To cut down or off with a vigorous swinging stroke. Now rare or Obsolete.	SWING
sweep, v.	1920	Cricket. To hit (the ball) with a sweep (sweep n. 5b). Also absol. or intransitive, to play a sweep.	SWING
sweep, v.	1577	To remove with a forcible continuous action; to brush off, away, aside.	BLOW STRIKE
sweep, v.	1560	Chiefly with away: To remove forcibly or as at one blow from its position or status, or out of existence; to do away with, destroy utterly.	BLOW STRIKE
sweep, v.	1635	a. To gather in or up, collect wholesale or at one stroke; esp. in to sweep the stakes (cf. sweepstake n.).	GATHER
sweep, v.	1942	U.S. To win every event in (a series of sporting events, etc.), or to take each of the main places in (a contest or event).	GATHER
sweep, v.	1616	To carry or trail along in a stately manner, as a flowing garment.	FLOWING MOTION

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

sweep, v.	date	Definition	key words
sweep, v.	1538	To pass over the surface of (something) in the manner of a broom or brush; to move over and in contact with; to brush, rub like (or as with) a brush.	BRUSH
sweep, v.	1892	To achieve widespread popularity throughout (a town, country, etc.). Also spec. in Politics, to gain control of by an overwhelming margin.	IMPETUS COVER
sweep, v.	1788	To range over (a region of sea or land), esp. to destroy, ravage, or capture; to scour. Also spec. with an aircraft as subject.	BLOW STRIKE WIPE OUT
sweep, v.	1638	To pass the fingers over the strings of a musical instrument so as to cause it to sound. (With the strings, or the instrument, as object.) Chiefly poetic.	BRUSH
sweep, v.	1744	To direct the eyes, or an optical instrument, to every part of (a region) in succession; to take a wide survey of, to survey or view in its whole extent, esp. with a glass or telescope. Also absol. or intransitive; in Astronomy to make systematic observations of a region of the heavens (cf. sweep n. 7).	SURVEY

Table 3: The senses of sweep [1300] in the OED

In column 2 the approximate date of attestation of the sense is given, in column 3 the OED definition used for analysis, in column 4 the key words in the definition, and in the column 5 the broader conceptual feature of the sense. As can be seen in the final column the senses of *sweep* tend to all trigger the same conceptual feature, with a few minor adaptations. There is no sign of major semantic shift from one feature to another.

2.1.2 Comparing *fl-* and *sw-* conceptual categories

There are 103 *fl-* monomorphemes, 180 senses in total which fit into 11 conceptual categories based on definition key words, and 270 combinations of features (see table 4). On the other hand, there are 217 *sw-* monomorphemes, and 330 senses out of 469 fit into the 18 conceptual categories based on definition key words. All words and senses were included if they were monomorphemic and consistent with recurring key words. A fair proportion (40%) carry labels such as obsolete, rare, and regional. As our purpose is to track change in behaviour rather than determine absolute behaviour, this methodology is considered suitable.

Feature abbr.	Conceptual categories	Examples
1 MTA	Move through air	*flap, flop, flick, flounce, flip, flit, fly, flee, flirt*
2 SV	Sudden violent	*flounce, flash, flit, flick*
3 FSC	Fail struggle confuse	*flop, flunk, flump, flummox,, flounder, flag (slacken), flivver*
4 SBT	Strike blow throw	*flick, flog, flail*
5 CJH	Clumsy jerky heavy (unsteady/awkward)	*fluster, flounder*
6 FLL	Flaccid limp loose	*flag, flop, flump*

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Feature abbr.	Conceptual categories	Examples
7 APF	Agitated panic fitful	*flurry, fluster, flicker*
8 MTL	Move through liquid (water)	*flash, flush, flow, flux, fleet, float, flask, flodder, flotter*
9 LDS	Light downy soft	*fluff, fleece, flake, floss, fleck*
10 DFF	Display flaunt flatter	*flatter, flutter, flare, flirt*
11 JS	Jeer sneer	*fleer, flout, flounce, flirt*

Table 4: Key words grouped into 11 conceptual categories for *fl-* words

Whereas there are 11 features for *fl-* words, there are 18 recurring features for *sw-* words in Table 5.

Feature abbr.	Feature	Examples
SSS	sway sweep swish	*Swimble, swabble, swaver*
SBS	strike blow swipe	*Switch, swash*
PSG	pressure swell grow	*Swivet, swench*
PSw	pressure swathe	*Swench, swaddle*
SSB	sway swagger boast	*Swell, swank, swagger*
CCA	compact cluster agitated	*swarm*
BF	big fellow	*Swad, swaddy*
FBW	flame burn waste	*Swither, swind, sweal*
DSS	deceive sway swindle	*Swike, swikel*
FSA	faint swoon agitated	*Sweer, swim, swarf*
CD	cool dark	*Swerk, swart, swale*
Dr	drink	*Swipe, swizzle, swoop, swill, swig*
Surf	surface	*Swarth, sward*
Hol	hollow	*Swire, swilly, swallow*
ExSw	exchange swap	*Switch, swap*
LTS	labour toil sweat	*Swat, sweat*
DD	deviate deflect	*Swerve, switch*
Sound	sound	*Swan, swear, swoosh*

Table 5: Key words grouped into 18 conceptual categories for *sw-* monomorphemes in the OED

Some of the features appear to show little relation to the more frequent features which are the 3 top lines in Table 5 (sway sweep swish, strike blow swipe, pressure grow swell). The green lines will be shown to be considered primary, whereas the grey lines correspond to what we have considered secondary features.

2.1.3 Combination of features, emergence of features and roots

The features themselves don't hold any specific meaning beyond the lexicographic cohesion of the definitions. However, it is possible to test the features based on four factors, 2 quan-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

titative factors and 2 qualitative factors. The quantitative factors in Table 6 are 1) the frequency of a feature and 2) the combinatorial properties of the feature. The qualitative factors are 3) the dates of emergence and post-emergence of a feature, and 4) the etymological roots of a feature. Results for *fl-* showed the existence of 3 primary features occurring frequently with strong form-meaning correlations: MTA, SBT, SV. In addition, there was some evidence of semantic shift towards secondary features such as FSC fail struggle confuse. The etymological roots of *fl-* are also consistent with a degree of **convergence** towards these primary features.

FEATURES	single feature	combination of 2	comb of 3	comb of 3 or more
Move through air	24	34	16	1
Sudden violent	7	20	12	2
fail struggle confuse	11	8	1	1
strike blow throw	9	12	8	1
clumsy jerky heavy	4	5	5	1
flaccid limp loose	3	7	1	0
agitated panic fitful	5	2	3	1
move through liquid	18	4	0	1
light downy soft	18	0	0	0
display flaunt flatter	7	11	0	0
jeer sneer	4	2	1	0
RAW TOTALS	110	105	47	9
FREQUENCY	40.74%	38.89%	17.41%	2.96%

Table 6: Combination of features for *fl-*

As opposed to *fl-*, the combination of features for *sw-* is very rare, contrary to the results for *fl-* words. The results for *sw-* showed that 3 primary features account for 57% of all senses: *sway*, *sweep swish*, *strike blow swipe*, and *pressure swell grow*. Categories with few tokens tend to be filled with the same lexeme and its senses, or even homonyms. Strikingly, the senses of *sw-* words (330 senses) appear to exhibit far less semantic shift between key conceptual features, indicating that *sw-* does not jump from one key conceptual feature to another to the same degree as *fl-* words. Polysemic words usually don't attract new conceptual features, like *swag* [1527] for instance. Sense 1 of *swag* [1527] which coincides with the first attested meaning activates sway sweep swish, and later senses don't activate new features coinciding with recurring key words.

Word	Date	OED definition	key words	feature 1
swag, v.	1527	To move unsteadily or heavily from side to side or up and down; to sway without control. of a pendulous part of the body, or of the whole person.	SWAY LURCH+ BIG CHUNKY	SWAY SWEEP SWISH
swag, v.	1611	of a structure or something erected or set in position, a boat, or the like. (Also occasionally of a rigid body, to get out of line.)	SWAY	SWAY SWEEP SWISH

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Word	Date	OED definition	key words	feature 1
swag, v.	1630	To sink down; to hang loosely or heavily; to sag. Also with down.	SINK	SINK
swag, v.	1846	To steal; to make away with (stolen property). Obsolete.	STEAL	STEAL
swag, v.	1861	To pack up (one's effects) in a 'swag'; to carry in a 'swag'; also, to wander about (the land) as a swagman.	WANDER	WANDER
swag, v.	1958	To push (a person) forcefully, to 'shove'; to take or snatch away roughly.	PUSH	PUSH

Table 7: Polysemy of *swag* [1527]

Incidentally, the correlation between roots of *sw-* also shows less consistency than for *fl-* words, where both Romance and Germanic roots tend to correlate with MTA (move through air).

2.2 Tracking change with dates of emergence

2.2.1 Evidence of shift; accounting for polysemy in the sense definitions

The OED definitions are central in identifying semantic change in the lexicon, the lexicographers rely on drawing analogical ties between etymological evidence and evidence of semantic shift. Awareness of anogical remotivation is sometimes explicit in the OED entries, although it is sometimes combined with a critical normative commentary ("contagion"). As Brewer (2016, p. 491) underlines: "it is clear that in a small number of instances both Murray and his fellow-*OED* lexicographers sought to impose their own views on the impropriety or undesirability of certain usages, even when these were amply attested by their quotation."

> All this means that the labels and usage notes in the first edition of the *OED* are a fine tool to identify and discriminate culturally significant vocabulary on the one hand, and/or lexicographical attitudes towards such vocabulary on the other, especially in view of *OED's* seminal status in English lexicography and its role as a cultural icon. (Brewer 2016, p. 493)

2.2.2 Do phonesthemes drive change?

The senses of *fl-* and *sw-* words were tracked using the feature analysis detailed above. Co-emergence of a feature with the first attestation is assumed to correlate with original etymological senses. Post-emergence of a new feature is assumed to correlate with semantic shift towards the new feature. Do phonesthemes affect semantic change? The data for *fl-* words in the OED in Table 8 provide some evidence of a shift of features.

	Sense 1	Feature 1	Sense 2	Feature 2
fleer [1400]	[1440] to flatter	DFF	[1549] to sneer, mock	JS
flounce [1542]	[1542] agitated, clumsy, violent	MTA and SV	[1751] n. a quick movement of the body expressing impartience or disdain	JS

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

	Sense 1	Feature 1	Sense 2	Feature 2
flourish [1303]	[1384] to display ostentatiously, to brandish	MTA	[1674] boast, brag, swagger	DFF
flummox [1837]	To bring to confusion, to do for, cause to fail, to confound	FSC	[1839] U.S. to give in, give up, collapse	FSC
flummer [1563]	To mumble, speak indistinctly	NONE	[1674] to deceive through flattery	DFF

Table 8: *Fl-* words exhibiting new senses

Although a shift has taken place, it is difficult to attribute the shift to the phonestheme only. It is generally thought that pathways of change are combined pathways rather than a single pathway. It is the combinatorial tendencies which may affect semantic change via analogical associations. However, compared to *fl-*, *sw-* words show hardly any shift, which seems to suggest that *fl-* and *sw-* have different behaviours.

We now look at combinatorial corpus patterns of a selection of *fl-* and *sw-* words in the OEC. The objective is to assess if behaviours do diverge, providing evidence of a different organising function.

3. Collocational behaviour in the OEC and phraseological patterns

3.1 Protocol and corpus collexeme analysis

3.1.1 Protocol and collexeme analysis in the OEC

The previous section showed that 1) *sw-* and *fl-* words do not share the same cohesion in the OED definition, and that 2) *sw-* words appear to undergo less semantic change than *fl-* words.

In this section I will use a collexeme analysis to track the behaviour of a sample of *sw-* words and *fl-* words in the Oxford English Corpus, using Sketch Engine (Kilgarrif et al. 2004). The driving question behind this test is to determine how collocational behaviour is able to track the semantic "cohesion" (the meaning of a phonestheme). To do this we ask**:** is there evidence of phonesthemic or sound symbolic clustering in the synonym set, and in the collocational context? The visual thesaurus function of Sketch Engine provides the candidates with the most similar lexicogrammatical collexemes. This provides a set of candidates for semantic proximity, The Sketch function of Sketch Engine provides the lexicogrammatical patterns of use of the target word in the corpus. Both of these tools help to determine paradigmatic semantic clustering as well as syntagmatic clustering. Using both tools should give evidence of the pressure or attraction of phonesthemic senses. This will be a preliminary test verifying if a small sample of common *fl-* and *sw-* words have different behaviours.

3.1.2 The OEC corpus and selection of lexemes

The Oxford English Corpus is a very large contemporary corpus of 2,1 billion words, combining spoken, written and computer-mediated (CMC) discourse. We know that phones-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

themic senses are triggered by surrounding context and type of discourse, so we do expect to find phraseological convergence. We have selected a group of central and frequent *sw-* and *fl-* words to test the results comparatively. We select *sweep*, *swish* and *sway* in the *sw-* category, and *fling*, *flip*, *flounce* for *fl*-words. The selected words are all essentially verbs (and nouns, but we are considering the verbs), and have varying levels of frequency (Table 9). Frequency affects entrenchment and therefore storage pathways in the mental lexicon. The more frequent a lexeme, the more likely it is to be selectively preferred by a speaker (Bybee 2013).

Word	Sweep	swish	Sway	Fling	Flip	Flounce
Frequency in OEC	60,000	1,215	13,325	10,119	19,446	535

Table 9: Compared frequencies of 6 target verbs in the OEC

3.2 Results for sway sweep and swish

3.2.1 sweep

Figure 1 shows the thesaurus of the highly frequent verb *sweep* followed by collexemes of *sweep* in lexicogrammatical positions in the OEC. The synonym candidates are based on the similarity of collocational behaviour and compared frequency: as figure 1 shows *sweep* is associated mainly with the metaphorical sense of *hit*, *blow*, *destroy*, *push*, *spread*. There are however few very close candidates, indicating that *sweep* appears to play a central role in the lexicon (as a potential foundation word).

Fig. 1: Thesaurus of sweep in the OEC

Sweep seems to have the most collocational behaviours, as the most frequent and widespread of the *sw-* words. The collexeme patterns in Table 10 show a number of specific subpatterns in contemporary discourse, which all trigger the same conceptual feature of *sw-*, whether used in its primary physical sense (*scrub*, *clean*, *mop*, *dust*) as in (2), or in its figurative sense (*sweep the board*, *sweep America*) as in (1).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Sweep								
and/or	**Freq**	**score**	**V* obj N**	**Freq**	**score**	**X* mod N**	**Freq**	**score**
mop	44	10.8	floor	495	8.6	generalization	392	9.5
sweep an mop			board	379	7.9	sweeping generalizations		
vacuum	26	9.9	swept the board			overhaul	165	7.9
dust	28	9.2	nation	345	7.5	a sweeping overhaul of the		
swept and dusted			sweeping the nation			vista	111	7.6
garnish	10	8.7	street	276	7.4	sweeping vistas of		
swept an garnished			sweep the streets			staircase	88	7.1
clean	59	8.2	Europe	138	7.1	a sweeping staircase		
sweeping and cleaning			sweeping Europe			reform	546	7
rake	8	8.2	globe	88	6.9	sweeping reforms		
scrub	9	7.9	sweeping the globe			change	1,369	6.9
sweep	8	7.6	country	651	6.9	sweeping changes		
wash	19	7.1	swept the countrys			move	275	6.7
sweeping and washing			fire	247	6.7	a sweeping move		
ignore	16	6.3	fire swept through the			bend	62	6.7
ignored or swept			region	156	6.5	sweeping bends		
damage	9	5.8	death as famine sweeps the region are a wake-up			panorama	52	6.7
pick	7	5.4	world	441	6.5	a sweeping panorama of		
			swept the world			driveway	51	6.6
			America	131	6.5	a sweeping driveway		
			swept Amerika			statement	422	6.4
			East	70	6.5	sweeping statements		
			sweeping the Middle East			epic	57	6.4
						a sweeping epic		

Table 10: Collexemes for sweep in the OEC

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

(1) Still, the growth likely would have been even bigger had it not been for the low-carbohydrate diet phenomenon that has been **sweeping** the country. (*Dairy Field*, May 2004)

(2) Over time I have helped design and mount exhibits, built exhibit furniture, given tours, taught school classes, designed and implemented a simple financial reporting system, vacuumed and **swept** the floors, fixed the plumbing, changed light bulbs, and even served as acting Director when the real Director was off having a baby. (*The Chronicle of the Early American Industries Association*, September 2001)

3.2.2 swish

Swish is the least frequent of the three *sw-* lexemes selected with 1,215 tokens. The thesaurus of *swish* shown in figure 2 shows *swish* to have far more candidates exhibiting the same collocational behaviour. The candidates are also all cohesive in terms of sound symbolic features, with most having some ties to phonesthemes or sound symbolic elements (*swirl*, *swish*, *flutter*, *flap*, *slosh*, *wag*, *whirl*).

Fig. 2: Thesaurus of *swish* in the OEC

The collexemes of *swish* seem to correlate with this higher propensity for sound-meaning association (see table 11). *Swish* is used in tandem with other *sw-* verbs *swipe* and *sway* which have the highest co-occurrence scores.

swish								
and/or	**Freq**	**score**	**N subj V***	**Freq**	**score**	**V* obj N**	**Freq**	**score**
swipe	3	9.5	tail	33	8.8	tail	46	7.6
sway	10	8.3	tail swishing			jumper	6	6.2
Rundgren while I swished and swayed an occassionally tilted			ponytail	3	8.1	throw	5	5.5
brush	5	7.9	skirt	8	7.9	skirt	5	5.1
spit	3	6.3	cape	3	7.9	bat	3	4.1
swallow	3	6.2	dress	6	6.7	past	8	3.7

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

swish								
and/or	**Freq**	**score**	**N subj V***	**Freq**	**score**	**V* obj N**	**Freq**	**score**
spin	3	5.7	surfer	4	6.4	blade	3	3.6
			hair	16	5.7	hair	8	2.4
			hair swishing			clothes	4	2.3
			Kirby	3	5.5	finger	3	1.2
			curtain	3	5.4	sound	3	1.2
			door	10	4.8	shot	4	0.9
			horse	3	3.2			
			money	4	2.2			

Table 11: Collexemes of *swish* in the OEC

Swish is always associated with sound correlations, either explicitly in (3) or implicitly as in (4):

(3) Bamboo – one of the most popular materials because its very lightweight. They also add an audible dimension with the **swishing** sound they produce. (*Alley: Home and Family articles*, 2005)

(4) But later in the season, after the flowers fade, grasses assume starring roles as their foliage turns shades of gold and red and their seed heads become kinetic sculptures, swaying and **swishing** in the breeze. (*Sunset Magazine*, October 2002)

3.2.3 sway

With 13,325 tokens, the verb *sway* has a number of polysemic extensions that do not correlate with sound symbolic associations; instead candidates with similar collocational behaviour coincide with a metaphoric sense of *sway*, such as *intimidate*, *persuade*, *prompt*, *influence* (see fig. 3).

Fig. 3: Thesaurus of *sway* in the OEC

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

sway								
and/or	**Freq**	**score**	**V* obj N**	**Freq**	**score**	**N subj V***	**Freq**	**score**
swing	30	9.1	hip	74	8.1	hip	45	8.5
swinging and swaying			voter	266	8.1	hips swaying		
bob	18	8.7	to sway voters			argument	101	7.6
bobbed and swayed			palm	39	7.5	opinion	43	7
rock	24	8.6	swaying palms and			swayed by public opinion		
rocked and swayed			opinion	330	7.3	consideration	28	6.9
roar	17	8.5	to sway public opinion			tree	59	6.9
robotic giants that sway and roar · Deinonychus Dash			jury	55	6.5	trees sway		
stagger	14	8.4	sway the jury			emotion	25	6.7
clap	28	8.4	scepter	10	6.3	swayed by emotion		
creak	13	8.3	undecided	9	6.2	prejudice	13	6.6
swish	10	8.3	Vote	148	6	they are swayed by prejudice, rely on		
Rundgren while I swished and swayed and occassionally tilted			juror	15	5.9	promise	15	6.5
bounce	17	8.3	lawmaker	17	5.8	rhetoric	16	6.5
bend	27	8	to sway lawmakers			be swayed by rhetoric		
bend and sway			electorate	13	5.8	grass	14	6.4
lurch	9	8	sway the electorate			palm	10	6.3
jiggle	8	7.9	judge	50	5.7	sentiment	13	6.2

Table 12: Collexemes of *sway* in the OEC

The collexemes in Table 12 show that in addition to the metaphorical sense of pressure, *sway* is associated with erratic manners of walking like *stagger* in (5) and also *swish* or *jiggle* in (6):

(5) A witness heard the crash and spotted a man getting out of the car, staggering and **swaying**, and heading away from the scene of the accident. (*This Is Wiltshire news stories*, October 2004 editions)

(6) The jiggling and **swaying** of the cab along with the gasoline smell leaking through the vent was getting to me and, for a while, I thought I might be carsick. (*The Boston Review*, April-May 2002)

These senses are associated with a sound-symbolic clustering, as can been seen in verbs like *jiggle* or *bounce*.

3.3 Results for flip, flounce, fling

3.3.1 flip

As far as *fl-* words are concerned, they appear more cohesive from the OED analysis. The three lexemes selected have different frequencies, *Flip* having the highest 19,446 number of tokens in the OEC, followed by *fling* 10,119, and the relatively infrequent *flounce* 535 tokens. The thesaurus for *flip* in figure 4 shows that candidates exhibiting similar collocational behaviour are quite far off at the periphery, similarly to *sweep*.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

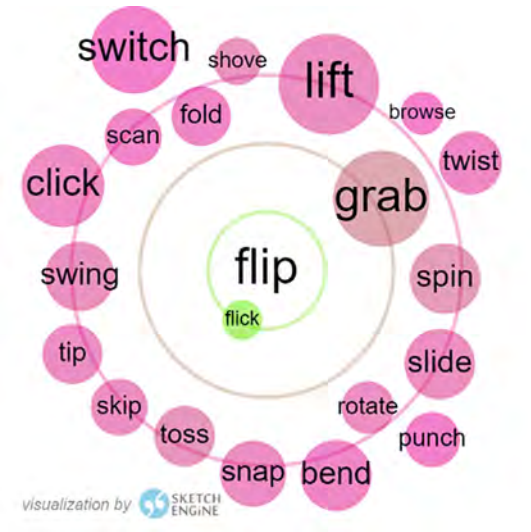


Fig. 4: Thesaurus for *flip*

The table of collexemes in table 13 shows that collexemes tend to be more cohesive with other phonesthemic or sound symbolic words (*flip*, *twist*, *spin*, *twist*) in the category *and/or*.

flip								
and/or	**Freq**	**score**	**V* obj N**	**Freq**	**score**	**N subj V***	**Freq**	**score**
flop	28	10.2	pancake	445	10.7	switch	20	7.3
flipped and flopped			posted by flipping pancades at			Oxide switch flips on . you		
rotate	27	9.3	switch	737	10.5	coin	16	7.2
rotate and flip			coin	492	10.2	coin flips		
flip	14	8.7	flip a coin			wig	9	6.7
flipped and flipped			burger	168	9.2	stomach	15	6.6
land	11	7.5	flipping burgers			stomach flipped		
flipped and landed			lid	93	7.9	calendar	8	6.2
spin	11	7.2	channel	180	7.8	back	14	6.2
toss	7	6.8	flipping channels			back flips		
twist	15	6.7	flop	47	7.7	boat	24	6.2
roll	11	6.7	flip flop			the boad flipped		
crash	8	6.7	page	283	7.2	seat	11	6
reverse	6	6.5	disc	82	6.9	car	69	5,9
jump	11	6.4	flip the disc for			car flipped over		
cook	11	6.2	bird	105	6.8	screen	10	5.5
flip an cook			flipping the bird			Kate	9	5.2
			wig	31	6.8	vehicle	16	5.2
			script	83	6.7	the vehicle flipped		
			flip the script					

Table 13: Collexemes of *flip* in the OEC

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

In the category Object of V, there are a number of specific subpatterns of *flip*; cooking metaphors (*flip pancakes*, *burgers*), *flip a coin*, and extended metaphors (*flipping channels*, *flip the script*, *flip the bird*, *flip your wig*).

(7) She thought she had gotten over her infatuation with him, because she didn't think of him as much anymore, though when she did, her stomach twisted and **flipped** inside. (Arurisonu, *A New*, 2004)[3]

(8) First up 'Hey Joe' made famous by Hendrix and reclaimed back to the Leaves, the Suzuki kids give it a complete lean and mean detox workout, drenched with retro keyboards the initially sombre rendition soon **flips** its wig to get down and dirty in fine style. (*Losing Today: Mark's Tales*, 2005)

3.3.2 fling

The thesaurus for *fling* in Figure 5 shows a number of lexemes with similar collocational behaviours, *toss*, *hurl*, *swing*, *grab dump*.

Fig. 5: Thesaurus of *fling* in the OEC

The collexeme analysis of *fling* shows a tendency towards a transitive pattern with an object realised by a noun referring to something undesirable and dysphemistic (*feces*, *excrement*, *insult*, *stones*), with more positive some subpatterns (*fling pillow*, *fling your arms*).

[3] https://www.fictionpress.com/s/1633584/1/A-New (last access: 11-04-2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

fling								
V* obj N	**Freq**	**score**	**N subj V***	**Freq**	**score**	**V* Part**	**Freq**	**score**
mud	32	7.4	Blast	9	6.2	aside	59	5
arm	349	7.3	Nick	5	4.7	flung aside		
flung her arms			And	5	4.5	together	8	4.1
feces	19	7.1	wave	6	4.3	around	70	3.4
pooh	17	7	arm	6	4.2	being flung around		
insult	30	6.7	wind	7	3.7	away	36	3.4
insults flung			opponent	5	3.5	flung away		
door	240	6.3	war	6	2.7	back	24	3.3
Amos	9	6.2	Darcy	5	2.6	flung back		
excrement	9	6.1	boy	6	2.4	about	47	2.9
cloak	11	5.9	girl	7	2.2	flung about		
spear	10	5.9	car	5	2.2	forward	25	2.8
pillow	12	5.8				when Amos was flung forward by the blast		
stone	37	5.7				off	97	2
flinging stones						flung off		
						down	105	1.9
						flung down		
						out	287	1.8
						flung out		
						up	114	0.2
						flung up		

Table 14: Collexemes of *fling* in the OEC

(9) Swan Dive, a figure we see hovering on tiptoe on a pedestal, his arms **flung** wide and his tie fluttering over his shoulder, may moments later land on his back like Poor Paul, who lies on the floor with one knee bent and his arms outstretched as if imploring. (*Art in America*, October 2001)

(10) If there were even a speck of dirt in the courtyard, if he **flung** a stone into the well or drummed a bit on the copper water pot, Valiamma would scold him unceasingly. (*The Little Magazine*, 2004)

3.3.3 flounce

Most of the synonym candidates in the thesaurus (see figure 8) are verbs of motion with similar frequencies (*scamper*, *totter*, *skulk*, *prance*, *waddle*) as well as *stomp* with a higher frequency. All candidates appear to have an expressive nature.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

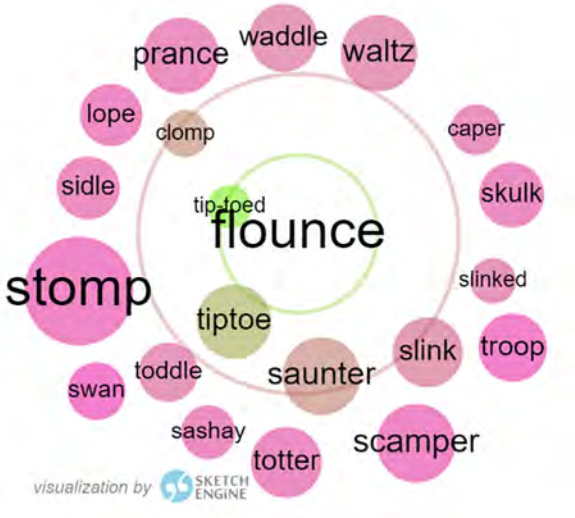


Fig. 6: Thesaurus of *flounce* in the OEC

The collexemes for *flounce* in Table 15 confirm the lack of data for the very small data set.

flounce								
V* obj N			**X* mod N**			**V* Part**		
noun	2	6.1	skirt	4	3.3	around	33	2.4
skirt	4	4.9	dress	7	2.3	flounce around		
						off	93	1.9
						flounced off		
						about	15	1.3
						flounce about		
						by	3	0.7

Table 15: Collexemes of the verb *flounce* in the OEC

4. Conclusion

This paper set out to answer a question; are phonesthemes evidence of a sublexical organising layer in the lexicon? We first compared the lexicographic behaviours of fl- and sw-monomorphemes in the OED. Then, a preliminary test based on collocational behaviour in a contemporary corpus showed that there is likely a divergence in behaviour based on the prevalence of phonosymbolic patterning, and also on the frequencies of usage of the lexemes in question. The highest frequency doesn't necessarily correlate with the greatest cohesion (*sweep* and *flip*),

Submorphemic layers are clearly present in the OED to varying degrees, and this variation in the acknowledgment of phonesthemes may be related to several factors, amongst which we can name two not necessarily conflicting phenomena:

1) the irregular treatment by OED lexicographers

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

2) the existence of distinct types of phonesthemes in the lexicon. These different types may themselves be tied to the words origins, as well as the frequency and salience of words carrying those phonesthemes in the lexicon.

We believe that further study focusing on the feedback between usage-based semantic analysis and usage-based lexicography is a worthwhile constructive way forward for improving the representation of the emergent nature of the historical lexicon, and for understanding the role of phonesthemes in the organisation of the lexicon.

References

Abramova, E./Fernandez, R. (2016): Questioning arbitrariness in language: a data-driven study of conventional iconicity. In: Proceedings of NAACL-HLT 2016, pp. 343–352. https://www.aclweb.org/anthology/N16-1038.pdf (last access: 19-03-2022).

Abramova, E./Fernandez, A./Sangati, F. (2013): Automatic labeling of phonesthemic senses. In: UC Merced (35), pp. 1696–1701.

Allan, K. (2012): Using OED as evidence. In: Allan, K./Robinson, J. (eds): Current methods in historical semantics. Berlin/Boston, pp. 17–40.

Atkins, B. T. S./Rundell, M. (2008): The Oxford guide to practical lexicography. Oxford.

Audring, J./Booij, G./Jackendoff, R. (2017): Menscheln, kibbelen, sparkle: Verbal diminutives between grammar and lexicon. In: Le Bruyn, B./Lestrade, S. (eds.): Linguistics in the Netherlands 2017. Amsterdam, pp. 1–15.

Bergen, B.-K. (2004): The psychological reality of phonaesthemes. In: Language 80 (2), pp. 290–311.

Bolinger, D. (1965): Forms of English: accent, morpheme, order. In: Abe, I./Tetsuya Kanekiyo, T. (eds): Cambridge/Tokyo, pp. 139–180.

Brewer, C. (2009): The Oxford English Dictionary's treatment of female-authored sources of the eighteenth century. In: Tieken-Boon van Ostade, I./Wan der Wurff, W. (eds.): Current issues in Late Modern English. Bern, pp. 209–238.

Brewer, C. (2016): Labelling and metalanguage. In: Durkin, P. (ed.): Oxford handbook of lexicography. Oxford, pp. 488–500.

Bybee, J. L. (2013): Usage-based theory and exemplar representations of constructions. In: Hoffmann, T./Trousdale, G. (eds.): The Oxford handbook of construction ggrammar. Oxford, pp. 1–14.

Considine, J. (2016): Historical dictionaries: history and development. Current issues. In: Durkin, P. (ed.): The Oxford handbook of lexicography. Oxford, pp. 163–175.

Durkin, P. (2016a): The Oxford handbook of lexicography. Oxford.

Durkin, P. (2016b): The OED and HTOED as tools in practical research: a test case examining the impact of loan words on areas of the core lexicon. In: Merja, K. (ed.): The Cambridge handbook of English historical linguistics. Cambridge, pp. 390–406.

Firth, J. (1930): Speech. London.

Flaksman, M. (2020): Pathways of de-iconization: how borrowing, semantic evolution and sound change obscure iconicity. In: Perniss, P./Fischer, O./Ljungberg, C. (eds.): Operationalizing iconicity. Amsterdam, pp. 75–104.

Geeraerts, D. (2016): Lexicography and theories of lexical semantics. In: Durkin, P. (ed.): The Oxford handbook of lexicography. Oxford, pp. 425–438.

Goldberg, A. (2006): Constructions at work: the nature of generalization in language. Oxford.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Hanks, P. (2013): Lexical analysis: norms and exploitations. Cambridge.

Kilgarriff, A./Rychlý, P./Smrž, P./Tugwell, D. (2004): The Sketch Engine. In: Information Technology. Lorient.

Liberman, A. (2008): An analytic dictionary of English etymology. Minneapolis/London.

Liberman, A. (2010a): Iconicity and etymology. In: Signergy, Iconicity in Language and Literature 9, pp. 243–258.

Liberman, A. (2010b): The state of English etymology (a few personal observations). In: Cloutier, R. A./ Hamilton-Brehm, A.-M./Kretzschmar, W. A., Jr. (eds.); Studies in the history of the English language V. Variation and change in English grammar and lexicon: contemporary approaches. Berlin/ New York, pp. 161–182.

Liberman, A. (2013): Flutes and flatterers. In: OUP blog July 10, 2013. https://blog.oup.com/2013/07/flute-word-origin-etymology/ (last access: 2022-04-08).

Mugglestone, L. (2009): The Oxford English Dictionary. In: Cowie, A. P. (ed.): The Oxford history of English lexicography. Oxford, pp. 230–259.

Ostermann, C. (2015): Cognitive lexicography. A new approach to lexicography making use of cognitive semantics. Berlin/Boston.

Otis, K./Sagi, E. (2008): Phonesthemes: a corpus-based analysis. In: Proceedings of the Annual Meeting of the Cognitive Science Society 30 (30), pp. 65–70.

Oxford English Dictionary: https://www.oed.com/ (last access: 2022-03-20).

Paton, B. (1995): New word lexicography and the OED. In: Dictionaries: Journal of the Dictionary Society of North America 16 (1995), pp. 79–89.

Pleyer, M./Hartmann S./Winters, J./Zlatev, J. (2017): Interaction and iconicity in the evolution of language. In: Interaction Studies 18 (3), pp. 303–313.

Sketch Engine: https://www.sketchengine.eu/ (last access: 2022-03-20).

Smith, C. A. (2016): Tracking semantic change in fl- monomorphemes in the OED. In: Journal of Historical Linguistics 6 (2), pp. 165–200.

Smith, C. A. (2018): Where do new words like boobage, flamage, ownage come from? Tracking the history of -age words from 1100 to 2000 in the OED3. In: Lexis 12. http://journals.openedition.org/lexis/2167, DOI: https://doi.org/10.4000/lexis.2167 (last access: 2022-04-11).

Smith, C. A. (2019): Approche cognitive diachronique de l'émergence du phonesthème fl- : réanalyse phonosymbolique et transmodalité dans le Oxford English Dictionary. In: Signifiances/Signifying 3 (1), pp. 36–62.

Smith, C. A. (2020): A case study of *-some* and *-able* derivatives in the OED3: examining the diachronic output and productivity of two competing adjectival suffixes. In: Lexis 16, http://journals.openedition.org/lexis/4793, DOI: https://doi.org/10.4000/lexis.4793 (last access: 2022-04-11).

Contact information

Chris A. Smith
University of Caen Normandy, CRISCO EA4255
chris.smith@unicaen.fr

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/ Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Data Models and Databases in Lexicography

EURALEX XX 22
XX EURALEX INTERNATIONAL CONGRESS
DICTIONARIES AND SOCIETY
12-16 July 2022, Mannheim, Germany

Thierry Declerck

INTEGRATION OF SIGN LANGUAGE LEXICAL DATA IN THE OntoLex-Lemon FRAMEWORK

Abstract We describe the status of work intending at including sign language lexical data within the OntoLex-Lemon framework. Our general goal is to provide for a multimodal extension to this framework, which was originally conceived for covering only the written and phonetic representation of lexical data. Our aim is to achieve in the longer term the same type of semantic interoperability between sign language lexical data as this is achieved for their spoken or written counterparts. We want also to achieve this goal across modalities: between sign language lexical data and spoken/written lexical data.

Keywords Sign Languages; OntoLex-Lemon; lexical data

1. Introduction

In the context of work dealing with the integration of multimodal lexical resources into the OntoLex-Lemon framework, which is described in (Cimiano/McCrae/Buitelaar 2016),[1] we investigate how to integrate lexical information included in Sign Language data. OntoLex-Lemon was originally covering only the written and phonetic representations of lexical data, as can be seen in the relation existing between the ontolex:LexicalEntry and ontolex:Form classes, which are displayed with the core module of OntoLex-Lemon in Figure 1.

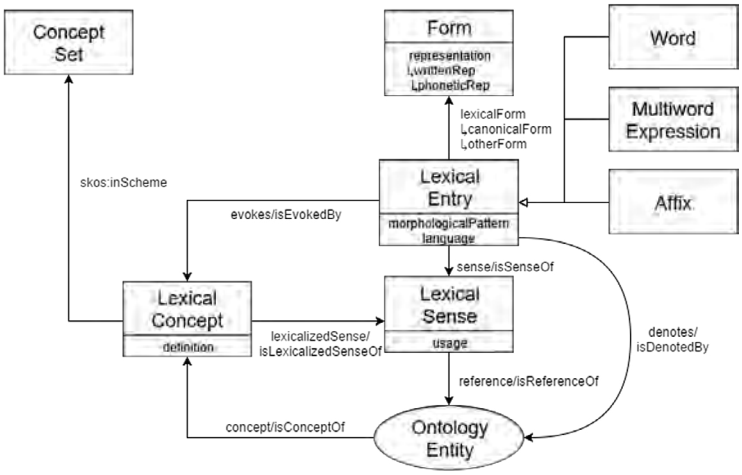


Fig. 1: Lemon_OntoLex_Core, taken from https://www.w3.org/2016/05/ontolex/

2. Consulted sources

We started our work by an extensive overview of the literature dedicated to the properties of sign languages (some of those works are included in the list of references), followed by a study of notational systems used for transcribing signs that mostly available in video or

[1] The full specification of OntoLex-Lemon is also available at https://www.w3.org/2016/05/ontolex/ (last access: 27-05-2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

pose streams. We concentrate in this paper on the possible representation of elements of such notational systems in the context of OntoLex-Lemon. Figure 2 gives a good overview of various ways of representing sign language data (here dealing with American Sign Language, taken from (Yin et al. 2021)), with three of them being notational transcriptions of the video or the pose streams: SignWriting,[2] HamNoSys[3] and Glosses.

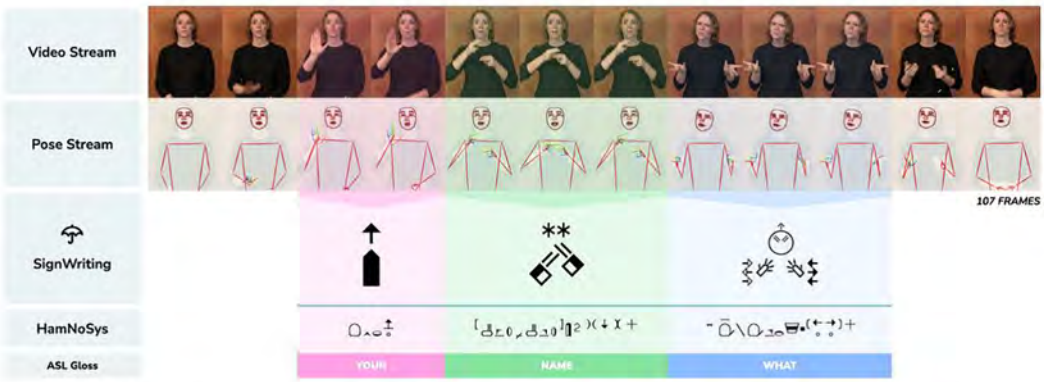


Figure 2: Various representations of American Sign Language. English translation: "What is your name?"

Fig. 2: Taken from Yin et al. (2021)

Glosses can be considered to label a sign (or a sequence of signs), as very often a corresponding (generally accepted) lexicon that could be used for annotating a sign (or a sequence of signs) is lacking. This issue is discussed in detail in (Ormel et al. 2010) and (Crasborn et al. 2012). If the glosses are to be seen more as labels used in the context of a corpus annotation process, it might make sense to consider their encoding within the "FrAC" OntoLex-Lemon extension module.[4]

The two other notational systems are representing (or transcribing) central elements of Sign Languages, like for example the shape and the orientation of the hands used by the signers, the interaction of the hands, their movements, also with respects to parts of the body and their activity, including repetitions, etc. For the time being we do not deal with the representation of facial elements, which left for a next stage of our work.

We focused for now on how to deal with the HamNoSys notational system, which breaks out a sign in four elements: handshape, orientation, location, and actions, as can be seen in Figure 3. But as HamNoSys per se is not machine-readable, we are making use of a conversion of it, called SiGML,[5] which is very often used as the input to avatar generation software. There exists a python implementation that transforms HamNoSys in SiGML, and which is

[2] More information about SignWriting can be found at https://en.wikipedia.org/wiki/SignWriting (last access: 27-05-2022).

[3] More information on HamNoSys can be found at https://www.sign-lang.uni-hamburg.de/dgs-korpus/files/inhalt_pdf/HamNoSys_2018.pdf (last access:27-05-2022). See also (Hanke 2004).

[4] "FrAC" stands "Frequency, Attestation and Corpus information", and is a potential extension module, that not only covers the requirements of digital lexicography, but also accommodates essential data structures for lexical information in natural language processing. See https://acoli-repo.github.io/ontolex-frac/ (last access:27-05-2022) for more detail.

[5] See https://vh.cmp.uea.ac.uk/index.php/SiGML (last access: 27-05-2022) for more details. See also Jennings et al. (2010):

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

described in Neves/Coheur/Nicolau (2020). The resulting notational code, which is displayed in Figure 4, is the one we use then to be included in OntoLex-Lemon, and from which we will able to link to a pose or video streaming object.

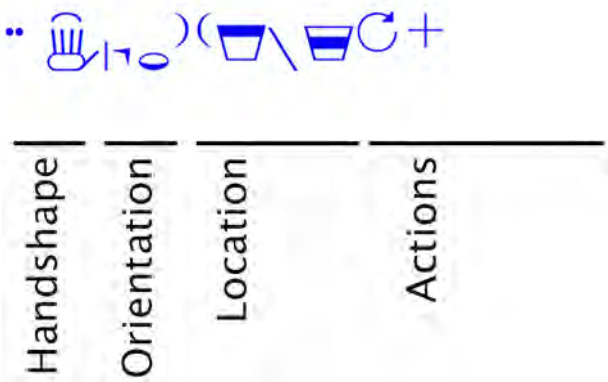


Fig. 3: The sign labelled with the German Word "Busch" in HamNoSys notation, using the four features: Handshape, Orientation, Location and Actions

```
(base) C:\Users\thde00\.spyder-py3\python_script\SignLanguage\HamNoSys2SiGML-
master\Original>python HamNoSys2SiGML.py "□□□□□□□□□□□□" (Busch)
<?xml version="1.0" encoding="UTF-8"?>
<sigml>
        <hns_sign gloss="(Busch)">
                <hamnosys_nonmanual/>
                <hamnosys_manual>
                        <hamsymmlr/>
                        <hamfinger2345/>
                        <hamfingerbendmod/>
                        <hamthumboutmod/>
                        <hamextfingeruo/>
                        <hampalmd/>
                        <hamclose/>
                        <hamshoulders/>
                        <hambetween/>
                        <hamchest/>
                        <hamcircleo/>
                        <hamrepeatfromstart/>
                </hamnosys_manual>
        </hns_sign>
</sigml>
```

Fig: 4: The SiGML conversion of the HamNoSys notation displayed in Figure 3, and which is used in our OntoLex-Lemon representation of sign language lexical data

3. Our current representation in OntoLex-Lemon

It is the kind of code displayed in Figure 4 that we can straightforwardly add to the OntoLex-Lemon framework, either introducing a new property to the ontolex:Form class (could be named ontolex:signRep) or by considering it as a written representation with a special tag "sigml", which is shown in Figure 5. In this example we can observe the complexity of the representation of such a sign, compared to the encoding for the written and phonetic representations. From this notational code we could link to video or pose streams that are displaying this sequence of signs.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

```
SignLangDatCat-2.ttl

ontolex:Busch_Sg
  rdf:type ontolex:Form ;
  lexinfo:gender lexinfo:masculine ;
  lexinfo:number lexinfo:singular ;
  rdfs:label "\"Busch\""@de ;
  ontolex:phoneticRep "\"buʃ\""@IPA ;
  ontolex:writtenRep "\"Busch\""@de ;
  ontolex:writtenRep "\"hamsymmlr
            hamfinger2345
            hamfingerbendmod
            hamthumboutmod
            hamextfingeruo
            hampalmd
            hamclose
            hamshoulders
            hambetween
            hamchest
            hamcircleo
            hamrepeatfromstart\""@de-DE-sigml ;
.
```

Fig. 5: Inclusion of the SiGML code within an instance of the ontolex:Form class, together with the encoding of the written and phonetic representations

We are currently investigating how the addition of this modality is affecting the representation and the linking its lexical data to lexical senses or lexical concepts. It might be that we need to duplicate lexical entries for being able to fully represent the contributions of sign language lexical data to meanings and concepts. As often stated, sign language is another type of natural language and its full representation (including semantics, etc.) might lead to a specific module extending OntoLex-Lemon. We also need to address the issue on how to represent cross-modal relations, as this was not needed in the case of the values of only the ontolex:writtenRep and ontolex:phonRep properties.

We are also working on establishing an ontology encoding all possible data categories associated with sign language (Declerck 2022). This ontology is re-using elements from the CLARIN concept repository (https://www.clarin.eu/content/clarin-concept-registry), the American Sign Language lexicon (https://asl-lex.org/visualization/), the British Sign Language dictionary (https://www.british-sign.co.uk/british-sign-language/dictionary/) as well as from Institute for German Sign Language and Communication of the Deaf at the University of Hamburg (https://www.idgs.uni-hamburg.de/). This ontology is also reusing elements of a former ontology for the Italian sign language, which is described in (Gennari/di Mascio 2007). Work will consist in linking the more than 250 constitutive elements of Sign Language included in this ontology to lexical descriptions represented in OntoLex-Lemon.

References

Cimiano, P./McCrae, J.-P./Buitelaar, P. (2016): Lexicon model for ontologies: final community report, 10 May 2016. https://www.w3.org/2016/05/ontolex/ (last access: 27-05-2022).

Crasborn, O./de Meijer, A. (2012). From corpus to lexicon: the creation of ID-glosses for the Corpus NGT. In: Crasborn, O. et al. (eds.): Proceedings of the 5th Workshop on the Representation and

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Processing of Sign Languages: Interactions between Corpus and Lexicon, Istanbul, Turkey, May. Istanbul: European Language Resources Association (ELRA), pp. 13–18.

Declerck, T. (2022): Towards a new ontology for sign languages. In: Proceedings of LREC 2022.

Gennari, R./di Mascio, T. (2007): An ontology for a web dictionary of Italian Sign Language. In: Proceedings of the Third International Conference on Web Information Systems and Technologies. Vol. 1: WEBIST, pp. 206–213.

Hanke, T. (2004): HamNoSys – representing sign language data in language resources and language processing contexts. In: Streiter, Oliver et al. (eds.): Proceedings of the Workshop on the Representation and Processing of Sign Languages: From SignWriting to Image Processing. Information techniques and their implications for teaching, documentation and communication, Lisbon, Portugal, May. Lisbon, pp. 1–6.

Jennings, V./Elliott, R./ Kennaway, R./Glauert, J. (2010): Requirements for a signing avatar. In: Hanke, T. (ed.): 4 th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies. Valletta, Malta, 22–23 May 2010. Valletta, pp. 133–136.

Neves, C./Coheur, L./Nicolau, H. (2020): HamNoSyS2SiGML: Translating HamNoSys Into SiGML. In: Proceedings of the 12th Language Resources and Evaluation Conference. Marseille, pp. 6035–6039.

Ormel, E./Crasborn, O./van der Kooij, E./van Dijken, L./Nauta, E. Y./Forster, J./Stein, D. (2010): Glossing a multi-purpose sign language corpus. In: Dreuw, Philippe et al. (eds.): Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies, Valletta, Malta. Valletta, pp. 186–191.

Quer, J./Cecchetto, C./Donati, C./Geraci, C./Kelepir, M./Pfau, R./Steinbach, M. (eds.) (2017): SignGram blueprint: a guide to sign language grammar writing. Publication resulting from the SignGram COST Action. https://parles.upf.edu/llocs/cost-signgram/node/18 (last access: 27-05-2022).

Yin, K./Moryossef, A./Hochgesang, J./Goldberg, J./Alikhani, M. (2021): Including signed languages in natural language processing. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, August 1–6, 2021. Vol. 1: Long papers, pp. 7347–7360.

Contact information

Thierry Declerck
DFKI GmbH, Multilinguality and Language Technology, Saarland Informatics Campus D3 2, Saarland, Germany
declerck@dfki.de

Acknowledgements

This paper is based upon work from the COST Action NexusLinguarum – European network for Web-centered linguistic data science (CA18209), supported by COST (European Cooperation in Science and Technology). It is also supported by the Horizon 2020 research and innovation programme with the projects Prêt-à-LLOD (grant agreement no. 825182) and ELEXIS (grant agreement no. 731015). We would like to thank the anonymous reviewers for their helpful comments.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Birgit Füreder

ÜBERLEGUNGEN ZUR MODELLIERUNG EINES MULTILINGUALEN ‚PERIPHRASTIKONS':

Ein französisch-italienisch-spanisch-englisch-deutscher Versuch

Abstract In the course of the last years, digital lexicography has opened up a variety of avenues fostering the conceptualisation, application and use of constructicons, a type of lexicographical reference work which has revealed itself highly promising in terms of connectivity and flexibility, at the same time, however, also challenging as to its technical implementation. The present paper takes up the ambitious aim to propose some reflections as well as a first draft for a possible model of a multilingual 'periphrasticon' as a subtype of a bigger constructicon focusing on a specific typology-related structural feature, i. e. periphrasticity. Taking periphrastic verbal constructions in French, Italian and Spanish as a starting point, it tries to sketch out a unified constructional network including not only equivalent (or corresponding) constructions within Romance, but also establishing (formal and functional) cross-linguistic connections to German and English. Comprising the major languages available to most language learners in (at least) German-speaking environments, the model is also supposed to pave the way for multilingual constructicography which, on the one hand, is able to account for intra- and cross-linguistic relations and, on the other hand, can also prove a valuable tool for language learning and use.

Keywords Multilingual lexicography; periphrastic constructions; French; Italian; Spanish; English; German

1. Ausgangspunkte und Zielsetzung oder: *Quo vadis, lexicographia*?

Der digitale Fortschritt der letzten Jahre und Jahrzehnte hat auch im Bereich der Lexikologie und Lexikographie eine Reihe an neuen und innovativen Wegen eröffnet. Dazu zählen nicht nur die vielfältigen Möglichkeiten der Zusammenstellung, Aufbereitung und Analyse groß angelegter Sprachkorpora, sondern auch neuartige Formen und Methoden der Systematisierung, Abbildung und Nutzbarmachung der Erkenntnisse über das Funktionieren und den Gebrauch von Sprache(n). Neben digital zugänglichen (und dementsprechend aufbereiteten) Wörterbüchern unterschiedlicher Art hat zuletzt im Besonderen die Konstruktikographie Fahrt aufgenommen, die sich um die Modellierung und Darstellung sprachlichen Wissens in Form unifizierter Netzwerke an Konstruktionen (unterschiedlicher Idiomatizität und Komplexität) bemüht. Erst die Multidimensionalität digitaler Ressourcen ermöglicht eine (zumindest ansatzweise) Implementierung dieser immensen Komplexität an Strukturen und Verbindungen, was in einem analogen Medium schlicht unmöglich wäre.

Ein vor allem in den romanischen Sprachen weit verbreiteter, lexikographisch bisher aber (vor allem sprachübergreifend) noch wenig aufgearbeiteter Konstruktionstyp scheint sich für eine derartige konstruktikographische Umsetzung in besonderem Maße zu eignen:

Periphrastische Konstruktionen zeichnen sich nicht nur durch höchst unterschiedliche Grade an Idiomatizität und Komplexität aus, sondern ebenso durch eine nicht außer Acht zu lassende sprachliche (und teils kulturelle) Spezifik, die auch den Erwerb besagter Konstruktionen maßgeblich beeinflussen kann. Nicht zuletzt erweisen sich derartige mehrteilige

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Wortverbindungen (allen voran Sprichwörter, idiomatische Wendungen, Kollokationen, aber auch Funktionsverbgefüge und Verbalperiphrasen) oft als Herausforderung im Fremdsprachenerwerb – umso mehr als auf interlinguale Vergleichbarkeit aufgrund anders strukturierter (oder nicht vorhandener) Äquivalente als Transferbrücke nur bedingt zurückgegriffen werden kann. Der vorliegende Beitrag setzt sich zum Ziel, ausgehend von einer Auswahl an romanischen Verbalperiphrasen Überlegungen zu einer holistischen, integrativen und zugleich sprachübergreifenden Modellierung dieses sprachlichen Bereichs in Form eines ‚Periphrastikons‘ anzustellen, das periphrastische Konstruktionen zum einen unter neuem theoretischen Licht präsentiert, zum anderen aber auch praxisorientierte Ansätze für eine mögliche Anwendung (nicht nur) in Spracherwerbskontexten versucht.

2. *Lexikon – Konstruktikon – Periphrastikon*? Eine terminologische Standortbestimmung

Bevor auf die beispielhafte Modellierung eines multilingualen Periphrastikons eingegangen wird, ist es zunächst notwendig, die Begriffe *Lexikon*, *Konstruktikon* und *Periphrastikon* terminologisch abzugrenzen. Der wahrscheinlich geläufigste Begriff, das *Lexikon*, kann auf verschiedene Sachverhalte referieren: 1) ein (physisches oder digitales) Wörterbuch aller Art; 2) den Teil der Sprache, der nach traditioneller und modularer Sprachbeschreibung komplementär zur *Grammatik* steht und den Bereich der *Wörter* umfasst, aus dem mithilfe des grammatikalischen Regelwerks Sätze geformt werden; und 3) die individuelle mentale Abbildung des in 2) beschriebenen Ausschnitts von Sprache, die – je nach theoretischer Ausrichtung – modular als binäre Einheit neben der *Grammatik* oder holistisch als Netzwerk kleinerer und größerer Spracheinheiten unterschiedlicher Natur modelliert wird. Vor allem aus konstruktivistischer Perspektive (gestützt durch Ergebnisse aus der psycholinguistischen Forschung kognitiver Prägung) erweist sich die Dichotomie zwischen Lexikon und Grammatik bisweilen als problematisch, was zur Annahme eines Kontinuums, das sich zwischen Lexikon und Grammatik[1] erstreckt und damit alle sprachlichen Phänomene in einem einheitlichen Ansatz zu beschreiben und erklären versucht, geführt hat (cf. u. a. Ziem/Lasch 2020, S. 90 ff. sowie Schafroth 2021).

Seit Anfang der 90er Jahre des vergangenen Jahrhunderts bereichert ein weiteres Konzept die Typologie der Nachschlagewerke: das sogenannte *Konstruktikon* (cf. u. a. Herbst 2016 und 2019). Analog zu der formseitig morphologischen Verschmelzung von *Konstruktion* und *Lexikon* gilt auch inhaltsseitig:

> Im Konstruktikon bilden grammatische Konstruktionen und lexikalische Elemente eine Einheit. Insofern das Konstruktikon also […] die Trennung von Lexikon und Grammatik aufhebt, lässt es sich näher bestimmen als ein taxonomisch strukturiertes Netzwerk form- und inhaltsseitig miteinander verbundener Konstruktionen, die sowohl hinsichtlich ihres Grades an Schematizität als auch hinsichtlich ihrer syntagmatischen Komplexität variieren […]. (Ziem 2014, S. 23)

Wie Herbst (2016, S. 170) erläutert, „kann die Entwicklung von Konstruktika durchaus in einer Reihe mit anderen Versuchen, bestimmte linguistische Theorien in Wörterbüchern quasi in die Praxis umzusetzen, gesehen werden“. Durch die wachsende Popularität der kognitiven Linguistik war die Forderung nach einer kognitiv orientierten Lexikographie

[1] Auch lexikogrammatisches Kontinuum oder Lexikon-Syntax-Kontinuum genannt.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

eine beinahe notwendige Konsequenz (cf. auch Ostermann 2015, S. 64–67). Entsprechende Projekte wurden und werden bereits für mehrere Sprachen erfolgreich umgesetzt – wie beispielsweise für das Deutsche im Rahmen der Projekte *FrameNet* und *Konstruktikon des Deutschen* an der Universität Düsseldorf (cf. Ziem et al. 2019–) sowie für das Englische an der Universität Erlangen-Nürnberg (cf. Herbst et al. 2016–).[2] Auch für die romanischen Sprachen ist derzeit ein größeres Projekt an der Universität Paderborn in Entwicklung (cf. Gévaudan et al. in Vorb.). Wie beim *Lexikon* kann der Begriff *Konstruktikon* sowohl auf ein physisches (oder digitales) Nachschlagewerk als auch auf eine mentale Entität referieren (cf. die Erläuterungen unter (1) und (3) oben). Zur Unterscheidung der beiden Bedeutungen schlägt Herbst (2016, S. 172) die Bezeichnungen *Referenzkonstruktikon* vs. *mentales Konstruktikon* vor und führt weiter aus:

> Ziel einer konstruktivistischen Lexikografie muss also die Schaffung eines Konstruktikons sein, das analog zum mentalen Konstruktikon Konstruktionen einer Sprache verzeichnet und für Benutzerinnen und Benutzer zugänglich macht […]. Dass Umfang und Beschreibungstiefe eines solchen Referenzkonstruktikons, wie bei traditionellen Wörterbüchern auch, von den Zielsetzungen bezüglich der Funktionen, die ein solches Konstruktikon erfüllen soll, und den intendierten Benutzergruppen abhängt, versteht sich von selbst.

In diesen Kontext ordnet sich auch das *Periphrastikon* als Teil bzw. Subtyp eines (Referenz-) Konstruktikons[3] ein. Ausgehend von einem bestimmten Konstruktionstyp (i. e. periphrastischen, i. e. S. verbalperiphrastischen Konstruktionen) sollen ausgewählte semantische Bereiche mit ihren jeweiligen Formen der Versprachlichung (zunächst einsprachig, in einem weiteren Schritt idealerweise mehrsprachig vernetzt) dargestellt werden. Eine derartige Herangehensweise ist freilich nicht für alle Sprachen (bzw. Sprachtypen) gleichermaßen relevant. Für Sprachen bzw. Sprachgruppen (wie beispielsweise die Familie der romanischen Sprachen), bei denen Periphrastizität eine zentrale Rolle spielt, kann eine solche unifizierte Darstellung (vor allem auch im Hinblick auf Lernende) wertvolle Dienste leisten – gerade auch, um strukturelle Gemeinsamkeiten und Unterschiede in der Versprachlichung bestimmter semantischer Konzepte zwischen den einzelnen Sprachen in einen größeren Kontext setzen und damit besser ersichtlich machen zu können.

3. Zur Modellierung eines multilingualen Periphrastikons

Analog zur Modellierung eines Konstruktikons kann mit Ziem (2014, S. 22) Folgendes auch für die Modellierung eines Periphrastikons (als Subtyp bzw. Teil Ersteren) festgehalten werden:

> Die Modellierung eines Konstruktikons zielt auf eine empirisch adäquate Beschreibung und Erklärung von Konstruktionen als form- und bedeutungsseitige Bestandteile eines ‚Netzwerkes‘, in welches sie so eingebettet sind, dass über Vererbungsbeziehungen alle form- und bedeutungsseitig lizenzierten Varianten (ab) gebildet werden können. […] Konstruktionen, also konventionalisierte Form-Be-

[2] Darüber hinaus existiert auch eine Reihe ähnlicher Projekte für das Schwedische, Japanische, Russische und Brasilianische Portugiesisch (cf. die entsprechenden Beiträge in Lyngfelt et al. (Hg.) 2018a).

[3] Um auch über die psychologische Validität eines solchen Konstrukts Aussagen treffen zu können, ist die Datenlage derzeit noch zu gering.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

> deutungspaare variierender Komplexität und Spezifität, haben alle gleichermaßen ihren Platz im Konstruktikon; sie bilden in ausdrucks- und inhaltsseitiger Hinsicht Knotenpunkte im Netzwerk.

Zu bedenken gilt es bei einem dermaßen umfangreich angelegten Unterfangen selbstverständlich, dass „[d]as Konstruktikon – im Sinne eines exhaustiven Modells grammatischer Strukturen einer natürlichen Sprache – […] in der empirischen Forschung immer nur ausschnitthaft und exemplarisch behandelt werden [kann]“ (ebd., S. 26).

Unter Berücksichtigung der erläuterten Schwierigkeiten und Einschränkungen soll im Folgenden eine mögliche Modellierung für ein mehrsprachiges Periphrastikon ausgehend von romanischen Verbalperiphrasen in Verbindung mit ihren deutschen und englischen Entsprechungen vorgestellt werden. Dies ist freilich ein recht ambitioniertes Vorhaben (vielleicht sogar „utopisch“, um es mit Herbsts Worten zu sagen, vgl. Herbst 2021, S. 26), betrachtet man zum einen die innerromanische Variation zwischen den einzelnen Sprachen und ihren jeweiligen Ausdrucksformen, zum anderen die adäquate Vernetzung mit Äquivalenten aus anderen Sprachen[4] (die form- und/oder bedeutungsseitig abweichen können) – ganz zu schweigen von der technischen Umsetzung eines derart komplexen Netzwerks an (teils unterschiedlichen) Konstruktionen.[5]

Zunächst ein kurzer Abriss zum konstruktionalen Typus *Verbalperiphrase*, der als Ausgangspunkt für den im vorliegenden Beitrag diskutierten Ausschnitt eines mehrsprachigen (auch für Lernzwecke geeigneten) Periphrastikons dient:

Verbalperiphrasen werden gemeinhin definiert als Verbindung von zwei (oder mehr) Verbalformen (eventuell verbunden durch eine Präposition oder Konjunktion), deren Bedeutung in vielen Fällen nicht kompositionell erschließbar ist. Aufgrund ihres gehäuften Vorkommens in romanischen Sprachen allgemein gelten sie als panromanisches Charakteristikum, wenngleich Inventar und Frequenz der betreffenden Konstruktionen zwischen den einzelnen Sprachen durchaus Unterschiede aufweisen. Diese verbalperiphrastischen Verbindungen dienen dem Ausdruck temporaler, aspektueller, modaler und diathetischer Werte, die im Deutschen teils durch ähnliche Konstruktionen, teils aber auch durch gänzlich anders strukturierte Äquivalente (wenn überhaupt) ausgedrückt werden. Dieser Umstand – in Kombination mit der unterschiedlichen sprachlichen Realisierung innerhalb der einzelnen romanischen Sprachen – resultiert für viele Lernende in einem unüberschaubaren, undurchsichtigen und unverständlichen ‚Wirrwarr‘. Eine Möglichkeit, dieses ‚Chaos‘ fassbarer und damit begreifbarer zu machen, könnte in der Erstellung eines Periphrastikons (als Teil eines größeren Konstruktikons) liegen, das sowohl mehrere romanische Sprachen (in einem ersten Schritt jedenfalls die Schulsprachen Französisch, Italienisch und Spanisch) als auch Deutsch und Englisch, die – zumindest in deutschsprachigen Erwerbskontexten – zu den festen Bestandteilen der meisten Sprachlernprofile gezählt werden können, umfasst.

[4] Bislang wurden Konstruktionen in erster Linie einzelsprachlich beschrieben, analysiert und dargestellt, in letzter Zeit erfreuen sich aber auch kontrastive Ansätze, die zwei oder mehr Sprachen berücksichtigen, immer größerer Beliebtheit (cf. z. B. Mellado Blanco/Mollica/Schafroth (Hg.) im Druck oder Benigni et al. 2015).

[5] Zu den (theoretisch-formalen und technisch-praktischen) Schwierigkeiten bei der Zusammenführung und Verbindung mehrerer einzelsprachlicher Konstruktika cf. u. a. Bäckström et al. (2014) sowie Lyng-felt et al. (2018b).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Exemplarisch sei dies nun kurz anhand des folgenden Beispiels illustriert:

Zum Ausdruck des semantischen Felds der INCHOATIVITÄT stehen in den romanischen Sprachen mehrere verbalperiphrastische Konstruktionen (mit verschiedenen Nuancen zur Art des Beginns der Handlung) zur Auswahl – z. B. fr. *commencer à/de* + Infinitiv, it. *cominciare/iniziare a* + Infinitiv, sp. *empezar/comenzar a* + Infinitiv (‚anfangen/beginnen‘) vs. fr. éclater de + Infinitiv, it. *scoppiare/sbottare a* + Infinitiv, sp. *echar(se)/romper a* + Infinitiv (‚plötzlich anfangen/beginnen‘). Während für die erste Gruppe an Periphrasen beispielsweise auch im Deutschen und Englischen ähnliche Konstruktionen zur Verfügung stehen (cf. z. B. dt. *anfangen/beginnen zu* + Infinitiv, engl. *start/begin to* + Infinitiv), sind die semantischen Äquivalente für die Konstruktionen der zweiten Gruppe häufig anders strukturiert: z. B. dt. *in Lachen/Weinen/Tränen/Jubel ausbrechen* (cf. auch engl. *burst into laughter/tears/flames*) oder *loslachen*, *losheulen*, *losschreien* etc.

Ein multilinguales Periphrastikon könnte dieser Komplexität zugleich auf mehreren Ebenen Rechnung tragen (sowohl formal als auch funktional) und somit – bei entsprechender digitaler Aufbereitung – auch als flexibles, multimediales Tool in Spracherwerbs- und Sprachverwendungskontexten aller Art hervorragend eingesetzt werden. Eine vereinfachte schematisierte Abbildung[6] könnte folgendermaßen aussehen:

[6] Folgende Abkürzungen werden verwendet:
Inf. = Infinitiv, Ger. = Gerundium; N = Nomen, V = Verb, Art. = Artikel.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

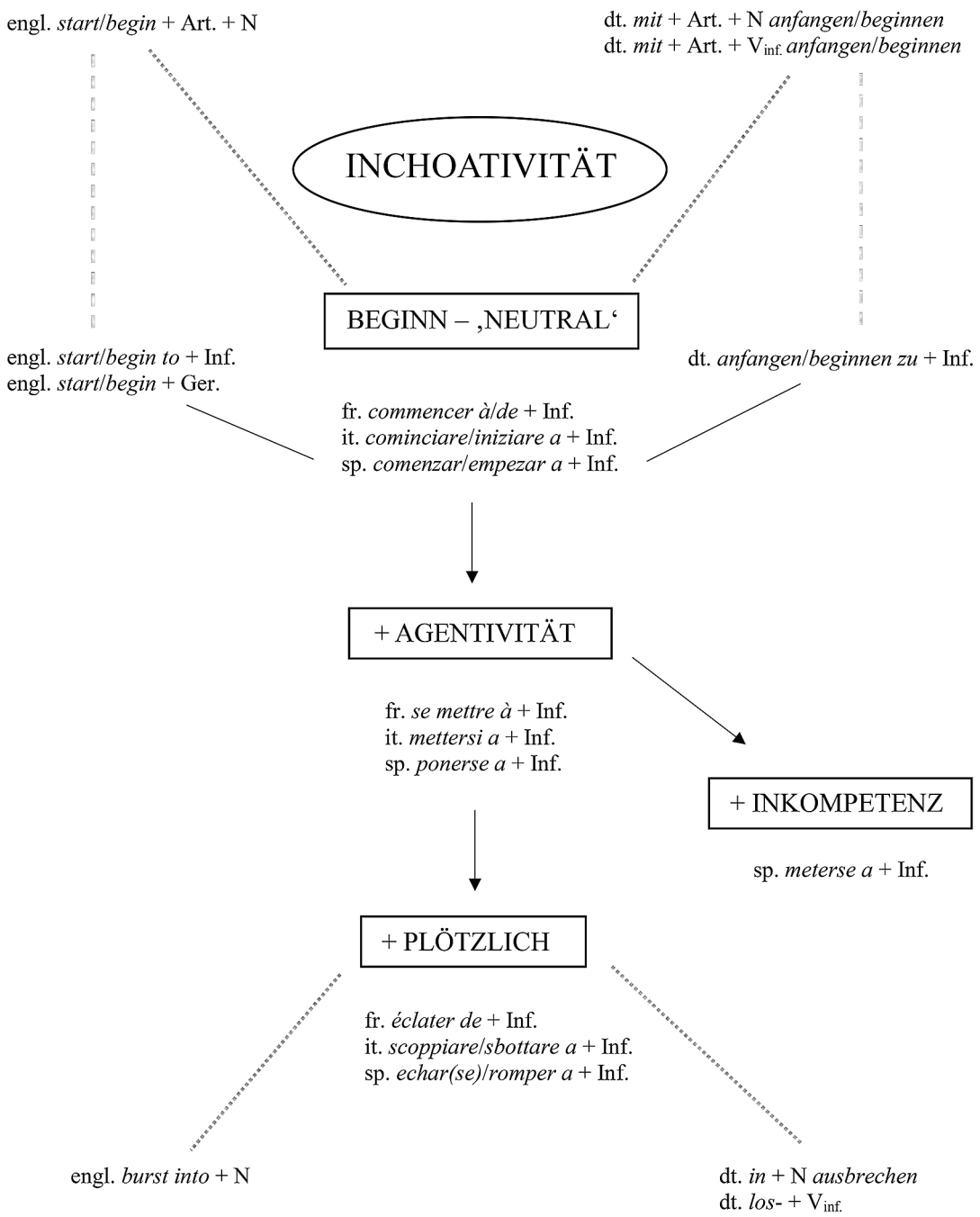


Abb. 1: Beispiel für einen Ausschnitt aus einem multilingualen Periphrastikon anhand des semantischen Felds der INCHOATIVITÄT

Neben den ‚Standard-Angaben‘, die lexikographische Werke üblicherweise bereitstellen (wie Informationen zu Aussprache, grammatikalischen Eigenschaften, Kollokationen, Variation im Gebrauch, Beispiel-Vorkommen etc., die entweder per *mouse-over*-Funktion oder per Mausklick angezeigt werden könnten), bietet eine digitale Umsetzung eines solchen Periphrastikons zudem die einzigartige Möglichkeit, mehrere Ebenen – nicht nur zwischen den einzelnen Sprachen, sondern auch innerhalb derselben – nach formalen und semantischen Kriterien miteinander zu vernetzen. So können einerseits formal ähnliche Konstruktionen miteinander in Verbindung gesetzt werden, andererseits aber auch semantische Verbindungen zwischen unterschiedlichen Konstruktionen deutlich gemacht werden (wie

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

beispielsweise verschiedene Ausdrucksformen der Referenz auf Zukünftiges cf. Herbst 2016, S. 196 für das Englische oder benachbarte semantische Bereiche cf. Füreder 2021, S. 166 am Beispiel aspektueller Verbalperiphrasen des Spanischen).[7]

Zur (naturgemäßen) Divergenz zwischen einem (zwar gebrauchsbasiert zusammengestellten, dennoch bis zu einem gewissen Grad abstrahierten und formalisierten) Referenzkonstruktikon und den individuellen mentalen Konstruktika der (potenziellen) Sprachbenutzer schreibt Herbst (2016, S. 172):

> Natürlich kann ein Referenzkonstruktikon keine 1:1-Abbildung eines mentalen Konstruktikons sein – schon deshalb nicht, weil man im Augenblick von empirisch fundierten Erkenntnissen über die Form eines solchen Konstruktikons im menschlichen Gehirn noch sehr weit entfernt ist und auch kein umfassendes Modell existiert, das allgemein akzeptiert wäre. Eine gewisse Parallele ergibt sich allerdings dadurch, dass das Referenzkonstruktikon idealerweise bottom up durch die Analyse von Korpusdaten[8] und entsprechende Abstraktionsprozesse durch die Lexikogrammatiker entsteht […]. Ein wesentlicher Unterschied besteht naturgemäß darin, dass Referenzkonstruktika langue-orientiert sein müssen und von daher die Unterschiedlichkeit der mentalen Konstruktika einzelner Sprecherinnen und Sprecher nicht abbilden können.

Nichtsdestotrotz kann ein solches (wenn auch abstrahiertes) visualisiertes Konstruktionsnetzwerk nicht nur wertvolle Dienste in Spracherwerb, -produktion und Translation leisten, sondern idealerweise zugleich auch ein Fenster zur Modellierung des mentalen Netzwerks periphrastischer Konstruktionen öffnen.[9]

4. Fazit und Perspektiven

Die Weiterentwicklung der Technik eröffnet nicht nur erweiterte Verwaltungs- und Zugriffsmöglichkeiten auf große Mengen an (Sprach-)Daten, sondern auch gänzlich neue Möglichkeiten der Systematisierung und Visualisierung komplexer (sprachlicher) Netzwerke. Dies kommt vor allem einer konstruktivistisch geprägten Analyse und Darstellung von Sprache entgegen, deren holistischer Ansatz die traditionelle Aufteilung von Wörterbuch und Grammatik hinfällig macht und dadurch auch für die Lexikographie – insbesondere für die Konstruktikographie – völlig neue Perspektiven schafft (cf. u. a. Herbst 2016, S. 170–172). Der vorliegende Beitrag hat sich zum Ziel gesetzt, die aktuellen konstruktikographischen Bestrebungen um einen Vorschlag für ein multilinguales Periphrastikon zu erweitern, das – ausgehend von den romanischen Sprachen – im Speziellen den Bereich periphrastischer Konstruktionen erörtert und mit dem Deutschen und Englischen verbindet. Die Vorzüge einer solchen Herangehensweise liegen zum einen in der Aufarbeitung eines sprachlichen Bereichs, der traditionell variabel zwischen Lexikon und Grammatik verortet wird (und dementsprechend uneinheitlich in Lexiko- bzw. Grammatikographie abgebildet ist);[10]

[7] Im Falle von Konstruktionen, deren Bestandteile (teil-)auxiliarisiert – und damit zumeist auch (teil-)desemantisiert sind – bieten derartige Quervernetzungen zudem die Möglichkeit, eine Verbindung zwischen der (voll-)lexikalischen Bedeutung und der (teil-)grammatikalisierten Bedeutung herzustellen und somit auch Grammatikalisierungs- bzw. Konstruktionalisierungsprozesse leichter nachvollziehbar zu machen.

[8] Für die romanischen (Schul-)Sprachen böten sich hierfür beispielsweise die Referenzkorpora CRFC für das Französische, CoLFIS für das Italienische und CORPES XXI für das Spanische an.

[9] Hierfür sind freilich auch einschlägige psycholinguistische Studien notwendig.

[10] Cf. in diesem Zusammenhang auch Schafroth (im Druck) zur digitalen Phraseologie.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

zum anderen in der Einbeziehung mehrerer (in diesem Fall romanischer und germanischer) Sprachen sowie deren konstruktionaler Verbindungen untereinander. Die digitale Umsetzung eines solch anspruchsvollen Projekts wird freilich kein Leichtes sein, der erhoffte Nutzen scheint jedoch vielversprechend.

Referenzen

Bäckström, L./Lyngfelt, B./Sköldberg, E. (2014): Towards interlingual constructicography: on correspondence between constructicon resources for English and Swedish. In: Constructions and Frames 6 (1), S. 9–33.

Benigni, V./Cotta Ramusino, P./Mollica, F./Schafroth, E. (2015): How to apply CxG to phraseology: a multilingual research project. In: Journal of Social Sciences 11 (3), S. 275–288.

CoLFIS = Corpus e Lessico di Frequenza dell'Italiano Scritto: Bertinetto, P. M./Burani, C./Laudanna, A./Marconi, L./Ratti, D./Rolando, C./Thornton, A. M. (1995): Corpus e Lessico di Frequenza dell'Italiano Scritto: CoLFIS. Scuola Normale Superiore di Pisa: Laboratorio di Linguistica 'Giovanni Nencioni'.

CORPES XXI = El Corpus del Español del Siglo XXI: Real Academia Española. https://www.rae.es/banco-de-datos/corpes-xxi (Stand: 20.3.2022).

CRFC = Corpus de référence du français contemporain = Siepmann, D./Bürgel, C./Diwersy, S. (2016): Le Corpus de référence du français contemporain (CRFC), un corpus massif du français largement diversifié par genres. In: SHS Web of Conferences 27, S. 1–13.

Füreder, B. (2021): Verbalperiphrasen aus konstruktionsgrammatischer Perspektive: Eine Annäherung. In: Döhla, H.-J./Hennemann, A. (Hg.): Konstruktionsgrammatische Zugänge zu romanischen Sprachen. Berlin, S. 143–170.

Gévaudan, P. et al. (in Vorb.): ConstructiCon – the Construction Lexicon of Romance Languages. https://kw.uni-paderborn.de/institut-fuer-romanistik/7 (Stand: 20.3.2022).

Herbst, T. (2016): Wörterbuch war gestern. Programm für ein unifiziertes Konstruktikon! In: Schierholz, S./Gouws, R./Hollós, Z./Wolski, W. (Hg.): Wörterbuchforschung und Lexikographie. Berlin/Boston, S. 169–206.

Herbst, T. (2019): Constructicons – a new type of reference work? In: Lexicographica 35, S. 3–14.

Herbst, T. (2021): Die Digitalisierung von Wortschatz und Grammatik – ein Konstruktikon für die Schule. In: Bürgel, C./Gévaudan, P./Siepmann, D. (Hg.): Sprachwissenschaft und Fremdsprachendidaktik: Konstruktionen und Konstruktionslernen. (= Stauffenburg Linguistik 119). Tübingen, S. 21–37.

Herbst, T. et al. (2016–): Englisches Konstruktikon. https://cris.fau.de/converis/portal/Project/125829976?auxfun=&lang=en_GB (Stand: 20.3.2022).

Lyngfelt, B./Borin, L./Ohara, K./Timponi Torrent, T. (Hg.) (2018a): Constructicography. Constructicon development across languages. (= Constructional Approaches to Language 22). Amsterdam/Philadelphia.

Lyngfelt, B./Timponi Torrent, T./Laviola, A./Bäckström, L./Hannesdóttir, A. H./da Silva Matos, E. E. (2018b): Aligning constructicons across languages: a trilingual comparison between English, Swedish, and Brazilian Portuguese. In: Lyngfelt, B./Borin, L./Ohara, K./Timponi Torrent, T. (Hg.): Constructicography. Constructicon development across languages. (= Constructional Approaches to Language 22). Amsterdam/Philadelphia, S. 255–302.

Mellado Blanco, C./Mollica, F./Schafroth, E. (Hg.) (im Druck): Konstruktionen zwischen Lexikon und Grammatik. Phrasem-Konstruktionen monolingual, bilingual und multilingual. (= Linguistik – Impulse & Tendenzen 101). Berlin/Boston.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Ostermann, C. (2015): Cognitive lexicography. A new approach to lexicography making use of cognitive semantics. Berlin/Boston.

Schafroth, E. (2021): Das Lexikon-Syntax-Kontinuum. In: Döhla, H.-J./Hennemann, A. (Hg.): Konstruktionsgrammatische Zugänge zu romanischen Sprachen. Berlin, S. 43–83.

Schafroth, E. (im Druck): Digitale Phraseologie. In: Becker, L./Kuhn, J./Ossenkop, C./Polzin-Haumann, C./Prifti, E. (Hg.): Digitale romanistische Sprachwissenschaft. Stand und Perspektiven. Tübingen.

Ziem, A. (2014): Konstruktionsgrammatische Konzepte eines Konstruktikons. In: Lasch, A./Ziem, A. (Hg.): Grammatik als Netzwerk von Konstruktionen. Sprachwissen im Fokus der Konstruktionsgrammatik. (= Sprache und Wissen 15). Berlin/New York, S. 15–34.

Ziem, A./Lasch, A. (2020): Konstruktionsgrammatik. Konzepte und Grundlagen gebrauchsbasierter Ansätze. 2. Auflage. (= Germanistische Arbeitshefte 44). Berlin/Boston.

Ziem, A. et al. (2019–): FrameNet und Konstruktikon des Deutschen. https://gsw.phil.hhu.de/ (Stand: 20.3.2022).

Kontaktinformationen

Birgit Füreder
Universität Salzburg
birgitursula.fuereder2@plus.ac.at

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

David Lindemann/Penny Labropoulou/Christiane Klaes

INTRODUCING LexMeta: A METADATA MODEL FOR LEXICAL RESOURCES

Abstract In this paper, we present **LexMeta**, a metadata model for the description of human-readable and computational lexical resources in catalogues. Our initial motivation is the extension of the **LexBib** knowledge graph with the addition of metadata for dictionaries, making it a catalogue *of* and *about* lexicographical works. The scope of the proposed model, however, is broader, aiming at the exchange of metadata with catalogues of Language Resources and Technologies and addressing a wider community of researchers besides lexicographers. For the definition of the LexMeta core classes and properties, we deploy widely used RDF vocabularies, mainly **Meta-Share**, a metadata model for Language Resources and Technologies, and **FRBR**, a model for bibliographic records.

Keywords Lexical resources metadata; linked data; Wikibase; semantic web

1. Introduction

In this paper we present LexMeta, a metadata model for the description of human-readable and computational lexical resources[1] in catalogues.

The goal is to develop a catalogue *of* and *about* lexicographical works to be integrated in the LexBib Wikibase Knowledge Graph of Lexicography and Dictionary Research, a research infrastructure targeting the lexicographic community. The LexBib project[2] (Lindemann/Kliche/Heid 2018; Kosem/Lindemann 2021) consists of various components among which LexBib Zotero[3] occupies a central place. This is a digital library of metalexicography research articles made available through the Zotero[4] platform, containing publicly available publication metadata, and a collection of full texts of articles available to the text processing objectives of the LexBib project.[5] It currently includes 10,000 metadata records for papers in several languages, out of which around 7,500 are included with their full texts. That bibliographical catalogue is represented as Linked Open Data (LOD) in LexBib Wikibase (Lindemann 2021) we present ongoing work concerning a workflow and software tool pipeline for collecting and curating bibliographical data of the domain of Lexicography and Dictionary Research, and data export in a custom JSON format as required by the Elexifinder application, a discovery portal for lexicographic literature. We present the employed software tools, which are all freely available and open source. A Wikibase instance has been chosen as central data repository. We also present requirements for bibliographical data to be suitable for import into Elexifinder; these include disambiguation of entities like natural persons and natural languages, and a processing of article full texts. Beyond the domain of Lexicography, the described workflow is applicable in general to single-domain small scale digital biblio-

[1] We use the terms "lexical resource" and "dictionary" interchangeably with a broad meaning, encompassing user dictionaries, general dictionaries, glossaries, thesauri, terminological lexica, etc.

[2] See https://lexbib.elex.is/wiki/Project:About.

[3] Accessible through https://lexbib.elex.is/wiki/LexBib_Zotero.

[4] Homepage at https://www.zotero.org/.

[5] For IPR reasons, we cannot make available physical copies of full text; nevertheless, where available, links to the locations where they can be accessed or downloaded from, are provided.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

graphies.”,”event”:”SiKDD 21 Slovenian KDD Conference, October 4th, 2021”,”event-place”:”Ljubljana”,”language”:”en”,”publisher-place”:”Ljubljana”,”title”:”Zotero to Elexifinder: Collection, curation, and migration of bibliographical data”,”URL”:”https://ailab.ijs.si/dunja/SiKDD2021/Papers/LindemannDavid.pdf”,”author”:[{“family”:”Lindemann”,”given”:”David”}],”issued”:{“date-parts”:[[“2021”,10,4]]}}}],”schema”:”https://github.com/citation-style-language/schema/raw/master/csl-citation.json”} . With the addition of metadata for dictionaries, the LexBib knowledge graph will cover lexicographical primary and secondary resources, along with other entity types related to both of these, such as persons, organisations, languages, places, events, and lexicographic terminology.

To increase the value and outreach of this catalogue, we foresee the import and export of metadata from and to other catalogues, especially those popular with our target audience. One such case is the CLARIN Virtual Language Observatory (VLO),[6] addressing researchers of the Social Sciences and Humanities disciplines. These catalogues serve different purposes and have, thus, adopted different approaches to the documentation of dictionaries: library catalogues of books mostly focus on bibliographical metadata, while catalogues of language resources, such as CLARIN, look at dictionaries (mainly those in digital form) as datasets and focus more on encoding information about their contents and accessing modes. Therefore, LexMeta seeks to bring together the metadata modelling approaches used in these two types of catalogues and cater for the description of lexical resources along both of these dimensions.

In the following sections, we present the background and main features of LexMeta, as well as its application in the LexBib catalogue. More specifically, section 2 presents the methodology for its development and gives an overview of the main models and deployed resources. section 3 describes the model itself illustrated with examples. section 4 introduces the current status of the LexBib catalogue of dictionaries and, finally, section 5 concludes with future plans.

2. Background

2.1 Requirements and methodology

The LexMeta model aims to cater for the description of lexical resources included in catalogues of libraries and repositories. It must satisfy the requirements and needs of the respective catalogue users but also have a broader outlook, considering recent developments and initiatives in the metadata and data-related areas, the most prominent being the formulation of the FAIR principles[7] (Wilkinson et al. 2016).

More specifically, in terms of content, the model must cover not only bibliographical metadata (e. g., title, author(s), publication date), but also information on the contents and accessibility of the resource, relations between versions of the same resource, and provenance metadata. It should also support easy discovery of the catalogue entries by both human users and machines, and thus exploit existing standards and best practices, especially those used by the involved communities. Extensibility and flexibility are important desiderata given the evolving data landscape. Interoperability with other schemas plays a crucial role in its design in order to facilitate exchange of metadata between catalogues.

[6] See https://vlo.clarin.eu.

[7] See https://www.go-fair.org/fair-principles/.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

For the design of the model, we first made an inventory of the metadata information that should be included in it based on the requirements of the envisaged use case. We also conducted a survey through which we identified a set of models and vocabularies that are popular in the targeted domains and explored their adoption for our needs as outlined in the next subsection.

2.2 Overview of models

For our survey we have investigated models and vocabularies used in the bibliographical and lexicographical domains and the domain of datasets. These are presented below with a short description of the features that are of interest for our model.

FRBR (Functional Requirements for Bibliographic Resources) is a conceptual model for describing bibliographic metadata (IFLA Study Group on the Functional Requirements for Bibliographic Records 1998)"language":"en","publisher-place":"Munich","title":"Functional Requirements for Bibliographic Records: Final Report.","URL":"http://www.ifla.org/en/publications/functional-requirements-for-bibliographic-records","author":[{"family":"IFLA Study Group on the Functional Requirements for Bibliographic Records","given":""}], "issued":{"date-parts":[["1998"]]}}}],"schema":"https://github.com/citation-style-language/schema/raw/master/csl-citation.json"} . It is an international standard implemented in numerous local applications. FRBR distinguishes between the concepts of *Work* (e. g., an abstract notion of a lexicographical creation), *Expression* (the realisation of a single work, such as a certain version or edition), and *Manifestation* (the distribution of a single realisation, e. g., on paper, or as a digital dataset) as core classes.

BIBO (The Bibliographic Ontology) was developed in the Semantic Web community, to provide a generic RDF vocabulary for describing bibliographic resources and citation relations. Building on widely used vocabularies such as Dublin Core,[8] BIBO provides specific classes and properties to classify and describe documents in a Linked Data environment. BIBO properties may relate to all FRBR core concepts.

The Meta-Share ontology **(MS-OWL** or **MS)**[9] (Gavrilidou et al. 2012; McCrae et al. 2015) caters for language resources, including data resources (structured or unstructured datasets, lexica, language models, etc.) and technologies used for language processing (taggers, parsers, machine translation applications, etc.). It builds around three key concepts: *resource type*, *media type* and *distribution*, which give rise to the core classes of the model. Focusing on lexical resources, the class ms:LexicalConceptualResource (subclass of ms:LanguageResource) covers resources such as term glossaries, dictionaries, semantic lexica, ontologies, etc., organised on the basis of lexical or conceptual units (lexical items, terms, concepts, phrases, etc.) along with supplementary information (e. g., grammatical, semantic, statistical information, etc.). The class ms:DatasetDistribution represents the accessible form of a resource, e. g., a spreadsheet or plain text file with the contents of a lexicon, or an online dictionary accessible through a user interface.[10] Properties are assigned to the most relevant class. Descriptive and administrative metadata, such as those used for identification purposes (title, description, etc.), recording provenance (creation, publication dates, creators, providers,

[8] See https://www.dublincore.org/specifications/dublin-core/dcmi-terms/.

[9] See http://w3id.org/meta-share/meta-share.

[10] MS includes an additional class for media parts not presented here because the LexMeta model is currently restricted to textual resources.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

etc.), are assigned to the class ms:LanguageResource, while more technical features and classification elements are attached to the appropriate subclasses. Thus, properties for ms:LexicalConceptualResource encode the subtype (e. g., computational lexicon, ontology, dictionary, etc.), and the contents of the resource (unit of description, types of accompanying linguistic and extralinguistic information, etc.). The ms:DatasetDistribution class provides information on how to access the resource (i. e., how and where it can be accessed), technical features of the physical files (such as size, format, character encoding) and licensing terms and conditions.

DCAT[11] (Data Catalog Vocabulary) is an RDF vocabulary for representing data catalogues. For our purposes, we have looked into two of its core classes and their properties.[12] dcat:Dataset represents "a collection of data, published or curated by a single agent or identifiable community; the notion of dataset is broad and inclusive, covering data in many forms, including numbers, text, pixels, imagery, sound and other multi-media, and potentially other types". A dictionary or any other lexical resource can safely be considered a dataset in DCAT terms. dcat:Distribution represents an accessible form of a dataset such as a downloadable file. The design of Meta-Share has been influenced by DCAT; thus, ms:LexicalConceptualResource and ms:DatasetDistribution are represented as subclasses of dcat:Dataset and dcat:Distribution respectively. Further alignments between them are currently under development.

The **LexVoc** Vocabulary of Lexicographic Terms[13] is part of the LexBib Wikibase graph. It is a structured controlled list of terms related to lexicographical and metalexicographical concepts. It has been developed by re-using and extending term lists from various authoritative sources and organising them in semantic domains with several goals in mind (Kosem/Lindemann 2021, section 3); among other applications, LexVoc terms are used for the content-describing indexation of LexBib bibliographical items, and can be used for the classification of dictionaries along various parameters. LexVoc is implemented using the SKOS model.[14]

2.3 Technical implementation considerations

With regard to the implementation of the model, we have decided to follow the Linked Data paradigm.[15] To this end, we have considered Semantic Web technologies (e. g., RDF, OWL,

[11] See https://www.w3.org/TR/vocab-dcat-3/

[12] The current version (v3), published in January 2022, as a working draft, is based around seven core classes. One of these, namely dcat:DatasetSeries, was introduced in this version. It also has a potential interest for the model and we are currently investigating its usefulness. This class stands for "a dataset that represents a collection of datasets published separately but sharing common characteristics that group them together".

[13] See http://lexbib.elex.is/wiki/LexVoc.

[14] The SKOS standard can be used for the representation of knowledge organization systems (KOS) such as thesauri, classification schemes, subject heading systems and taxonomies within the framework of the Semantic Web. For more information, see https://www.w3.org/TR/2009/REC-skos-reference-20090818/.

[15] For an introduction to Linked Data, see https://www.w3.org/standards/semanticweb/data.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

SKOS) and the model behind Wikidata,[16] an open knowledge graph based on the Wikibase software.[17]

LexBib Wikibase is an instance of Wikibase, an open source software solution. Wikimedia Germany,[18] a non-profit organisation, is in charge of providing Wikibase as a service to a broader community, by an endeavour called Wikibase Cloud.[19] The goal is to enable an ecosystem of federated Wikibases, with Wikidata as the central hub; federation is possible through shared persistent identifiers and an interoperable querying standard, SPARQL, which allows for accessing different Wikibases at the same time. A Wikibase may provide additional data describing entities represented on Wikidata.

Advantages of Wikibase compared to other Linked Open Data (LOD) database infrastructures are described in Lindemann (2021) we present ongoing work concerning a workflow and software tool pipeline for collecting and curating bibliographical data of the domain of Lexicography and Dictionary Research, and data export in a custom JSON format as required by the Elexifinder application, a discovery portal for lexicographic literature. We present the employed software tools, which are all freely available and open source. A Wikibase instance has been chosen as central data repository. We also present requirements for bibliographical data to be suitable for import into Elexifinder; these include disambiguation of entities like natural persons and natural languages, and a processing of article full texts. Beyond the domain of Lexicography, the described workflow is applicable in general to single-domain small scale digital bibliographies.”,”event”:”SiKDD 21 Slovenian KDD Conference, October 4th, 2021”,”event-place”:”Ljubljana”,”language”:”en”,”publisher-place”:”Ljubljana”,”title”:”Zotero to Elexifinder: Collection, curation, and migration of bibliographical data”,”URL”:”https://ailab.ijs.si/dunja/SiKDD2021/Papers/LindemannDavid.pdf”,”author”:[{“family”:”Lindemann”,”given”:”David”}],”issued”:{“date-parts”:[[“2021”,10,4]]}},”suppress-author”:true}],”schema”:”https://github.com/citation-style-language/schema/raw/master/csl-citation.json”} . Furthermore, Wikibase as an infrastructure supports FAIR data and metadata; (meta)data in Wikibase are

- findable by machines through unique and persistent identifiers,
- accessible using standardised protocols (in particular, SPARQL),
- interoperable through the use of broadly used vocabularies that follow the same FAIR principles, and allow cross-references to other datasets, and
- reusable through appropriate licensing.

3. LexMeta Model

3.1 LexMeta presentation

Through the analysis of the descriptive requirements for our model and the survey of the models and vocabularies, we have established alignments between them and identified conceptual gaps, for which we have introduced new elements in a unified model.

[16] See https://www.mediawiki.org/wiki/Wikibase/DataModel.

[17] See https://wikiba.se/.

[18] See http://wikimedia.de.

[19] See http://wikibase.cloud.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

The LexMeta model is built around three main classes,[20] which follow the FRBR and relevant MS conceptual distinctions:

- the *Lexicographic Work* (lwb:Q41)[21] corresponds to the abstract notion of a lexicographical creation and is defined as subclass of frbr:Work;[22]
- *Lexical/Conceptual Resource* (*LCR*, lwb:Q4) represents the realisation of a single work, such as a certain version or edition of a lexicographic work, and corresponds to frbr:Expression and ms:LexicalConceptualResource;
- *LCR Distribution* (lwb:Q24) is the physical form in which a lexicographical work is realized (e. g., as a printed book or as a digital file), and is aligned to frbr:Manifestation and ms:DatasetDistribution.

This distinction allows us to group and link different publications (e. g., print publications, reprints, and digital versions) with the same content as well as to describe them more consistently by attaching their properties at the appropriate level.

Lexicographic Work groups the various editions and versions (*expressions*/*LCRs*) of the same *work*. Content-describing metadata are common across *manifestations* (*distributions*) of the same *expression* (*LCR*) and are assigned to the *LCR* level. Publication metadata and technical features are attached at the *distribution* level.

More specifically, properties for a *Lexicographic Work* include identification metadata (title, identifier) and the *has realisation* property (lwb:Q118, frbr:realization) that links it to the *LCR* objects.

Properties attached to the *LCR* class relate to identification, administrative and provenance metadata (e. g., title,[23] author, holder of Intellectual Property Rights, etc.) that are common across all its Distributions. The property *has distribution* (lwb:P55) is used to link the *LCR* to one or more *LCR Distributions* while specific properties (taken from MS) are used to relate different *LCRs* to each other, e. g. *replaces LCR* (lwb:P135, ms:replaces). To encode the language(s) of the contents, four distinct properties are included: *source* and *target language* (for multilingual resources), *object language* and *metalanguage*. Properties describing LCR structure and type include the following:

- *lemma type* (lwb:P151), describing types of headwords included in a dictionary (e. g., single-word or multi-word units, abbreviations, neologisms, etc.),
- *linguality type* (lwb:P115), indicating whether the LCR describes one, two or more languages,
- *dictionary scope type* (lwb:P90), pointing to dictionary typology terms, such as "learner dictionary", "dialect dictionary", "etymological dictionary",

[20] Hereafter, we use the terms *class* and *property*, as in RDF vocabularies, to represent the objects we wish to describe and their features respectively.

[21] The namespace prefix "lwb" (short for "LexBib Wikibase" resolves to http://lexbib.elex.is/entity/).

[22] The MS ontology has no similar class; for the connection between versions of the same resource, it relies solely on properties that link them together (e. g. ms:isContinuedby, ms:isPartOf, etc.).

[23] Title (and other identification data) is a property that can be used for all the classes. This is deliberate to allow for cases where, for instance, distributions have different titles from that of the LCR and between them (e. g. "Paperback edition of Dictionary X", "Dictionary X: the online version", etc.).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

- *dictionary function type* (lwb:P120), pointing to basic terms describing communicative and cognitive dictionary functions, e. g. "text translation",
- *dictionary access type* (lwb:P121), with two values, "onomasiological" and "semasiological dictionary",
- *microstructure feature* (lwb:P127), pointing to terms describing microstructural data presentation features as well as linguistic features of the presented content,
- *dictionary text part* (lwb:P152), indicating parts present in the dictionary text, such as front and back matters, and types of entries.

At *LCR Distribution* we attach publication metadata (e. g., publication date, publisher, ISBN), as found in a library catalogue, compatible with how publication metadata are represented in LexBib for metalexicographical publications.[24] We also attach properties describing how they can be accessed, i. e., the form of access or distribution type (e. g., "dictionary book publication" or "dictionary app.") and the URL where they can be accessed or downloaded. Where possible, we have opted for SKOS controlled vocabularies instead of free text to increase consistency and standardisation. Re-use of existing vocabularies, such as LexVoc, is preferred. In some cases, we have imported and enriched the LexMeta vocabularies with terms from other vocabularies. For example, the vocabulary containing terms that describe dictionary microstructure features is an extension of the MS vocabulary of content types, which is used in the range of the property ms:linguisticInformation.

3.2 Implementation

For the implementation of the model, we have decided to use two conventions: (1), Following the Wikibase data model, as an ontology of Wikibase entities, since the catalogue will be integrated in the LexBib Wikibase, and (2), as an OWL ontology,[25] a widely used formal knowledge representation language for the description of digital data.

In a more detailed documentation,[26] we specify the LexMeta core classes as implemented in the LexBib Wikibase, and the LexMeta properties attached to each of these classes, their datatype, and, for properties that take values from controlled vocabularies or classes of items, the respective set of values. We also include the alignments to the classes and properties of the vocabularies presented in section 2.2.

In the LexBib Wikibase implementation, LexMeta classes and properties are represented using URIs from the LexBib Wikibase's own namespace, and following Wikibase naming conventions, i. e. numeral identifiers preceded by the letter Q for items (i. e. classes and instances), and the letter P for properties. The LexMeta ontology is currently under construction. Where possible, we opt for re-using classes and properties from other vocabularies (mainly Meta-Share, BIBO, DCterms, etc.) instead of creating new ones.

The alignment between the two forms is foreseen at both sides. At the LexMeta OWL side, the OWL equivalence semantic relations can be used for linking to the LexBib Wikibase entities. In the LexBib Wikibase, this is already represented with a property of type "exter-

[24] This allows, at the same time, the creation of bibliographic items for LCR distributions on LexBib Zotero in a straightforward way.

[25] See https://www.w3.org/OWL/.

[26] Accessible at http://lexbib.elex.is/wiki/LexMeta.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

nal identifier" (lwb:P42), which links to the identifier of the equivalent LexMeta OWL entities.[27] Hence, in a data export of the LexBib entries, the LexBib identifier can be translated to its LexMeta OWL equivalent.[28]

4. Population of LexBib

The LexBib catalogue is already populated with metadata of dictionaries; that was initially done for a set of example items which were manually created and annotated with properties from our model.

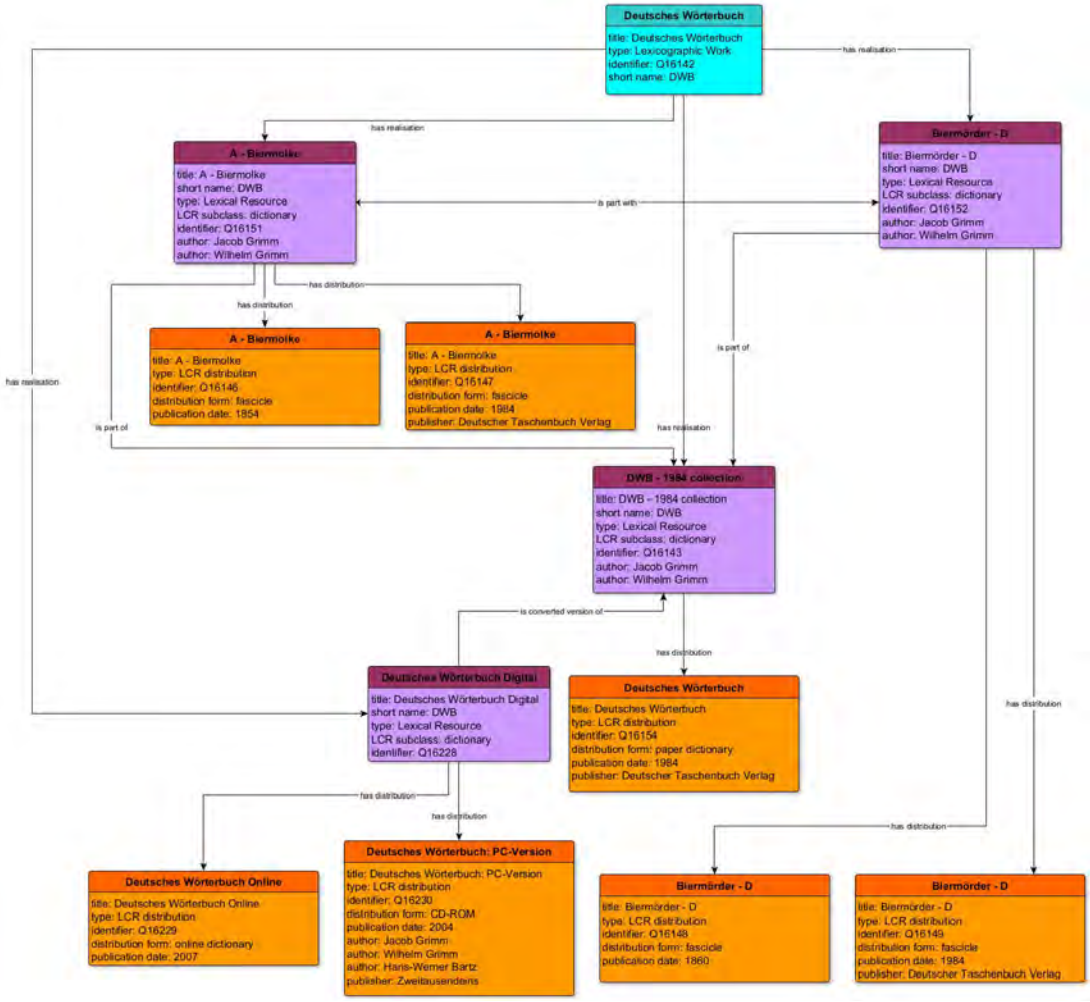


Fig. 1: Relations between instances of the LexMeta core classes

As an example that illustrates the complexity of relations our model allows between entities of the three core classes (see Fig. 1), *Deutsches Wörterbuch*, a lexicographical endeavour started in the mid 1850ies by the Grimm brothers, can be represented as one *Lexicographic*

[27] LexMeta OWL namespace is http://w3id.org/meta-share/lexmeta/.

[28] The Wikibase data model has also a particular data structure vis-a-vis the RDF model. Thus, the mapping process includes a step for the conversion of this structure, see http://lexbib.elex.is/wiki/LexMeta_OWL.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

work[29] with several expressions as *LCR*. The work was initially released as a set of fascicles, with different contributors and content features, and each had a different distribution: the book publications date from 1854 (fascicle 1) to 1954 (fascicle 32). After 1984, the fascicles are reprinted, with the same contents as the original ones, and are thus considered distributions of the original LCR. At the same time, they were issued as a complete collection, with a different size (i. e., the original LCR is linked to this with a part-of relation). This collection distributed in print in 1984 was later converted to a digital resource, which is distributed as an offline electronic dictionary, and also made accessible through a web portal.[30]

Other metadata entries already accessible at LexBib Wikibase stem from various source catalogues, such as OBELEX-dict,[31] Glottolog,[32] Worldcat,[33] and Wikidata. Properties that relate items describing dictionaries to items describing metalexicographical publications are part of LexMeta, namely *is reviewed in* (lwb:P26) and *cites* (lwb:P147, bibo:cites), which enables setting review and citation relations in the graph.

5. Conclusions and outlook

In this paper, we have presented the LexMeta model for lexical resources and its use in the population of the LexBib knowledge graph with metadata of dictionaries.

We are currently in discussions with scholars from the lexicographical and linguistic linked data communities in the framework of the ELEXIS[34] and NexusLinguarum[35] projects respectively and expect valuable feedback from them that will be used for the improvement of the model and its documentation. We are also collaborating on proposals aiming at a (community-driven) curation of lexicographical primary and secondary resources metadata on the LexBib wikibase, including assertions regarding review and citation relations.

Among our future plans is the enrichment of the LexBib catalogue with (mass) imports of metadata from other catalogues and, if and where needed, alignment of LexMeta with models used for these catalogues as well as with more general widespread metadata models for data resources.

In addition, the LexBib catalogue is planned to be made available through a CLARIN Knowledge Centre[36] (under construction) dedicated to Lexicography. In this case, the metadata of dictionaries will be exposed for harvesting by the CLARIN VLO, which is based on the OAI-PMH protocol and the use of metadata profiles that are compatible with the Component MetaData Infrastructure (CMDI) framework[37] (Broeder et al. 2012; International Organization for Standardization 2020). The conversion of the metadata into a CMDI-compatible profile can benefit from the fact that the Meta-Share schema is already included

[29] The URI of *Deutsches Wörterbuch*, an entity of class Work, is http://lexbib.elex.is/entity/Q16142.

[30] Graphical representations of these relations are available at https://lexbib.elex.is/wiki/Dictionaries.

[31] See https://www.owid.de/obelex/dict/en.

[32] See https://glottolog.org/langdoc, selecting “Doctype dictionary”.

[33] Accessible at https://www.worldcat.org/.

[34] Homepage at https://elex.is/.

[35] Homepage at https://nexuslinguarum.eu/.

[36] See https://www.clarin.eu/content/knowledge-centres.

[37] See https://www.clarin.eu/content/component-metadata.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

among them and can therefore be based on the re-use of the Meta-Share entities in the LexMeta Model.

References

Broeder, D. et al. (2012): CMDI: a component metadata infrastructure. In: Proceedings of the workshop describing language resources with metadata: towards flexibility and interoperability in the documentation of language resources. LREC 2012, May 22, 2012, Istanbul, Turkey. Paris, pp. 1–4. https://ids-pub.bsz-bw.de/frontdoor/index/index/docId/10867.

Gavrilidou, M. et al. (2012): The META-SHARE Metadata Schema for the description of language resources. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12). Istanbul, Turkey, pp. 1090–1097. http://www.lrec-conf.org/proceedings/lrec2012/pdf/998_Paper.pdf.

IFLA Study Group on the Functional Requirements for Bibliographic Records (1998): Functional requirements for bibliographic records: final report. Munich. http://www.ifla.org/en/publications/functional-requirements-for-bibliographic-records.

International Organization for Standardization (2020): ISO 24622-1:2015: Language resource management – Component Metadata Infrastructure (CMDI) – Part 1: The Component Metadata Model. ISO. https://www.iso.org/standard/37336.html.

Kosem, I./Lindemann, D. (2021): New developments in Elexifinder, a discovery portal for lexicographic literature. In: Gavriilidou, Z./Mitits, L./Kiosses, S. (eds.): Lexicography for Inclusion: Proceedings of the 19th EURALEX International Congress, 7–11 September 2021, Alexandroupolis, Vol. 2. Alexandroupolis, pp. 759–766. https://euralex2020.gr/proceedings-volume-2/.

Lindemann, D. (2021): Zotero to Elexifinder: collection, curation, and migration of bibliographical data. In: SiKDD 21 Slovenian KDD Conference, October 4th, 2021. Ljubljana. https://ailab.ijs.si/dunja/SiKDD2021/Papers/LindemannDavid.pdf.

Lindemann, D./Kliche, F./Heid, U. (2018): LexBib: a corpus and bibliography of netalexicographcal publications. In: Proceedings of EURALEX 2018. Ljubljana, pp. 699–712. http://euralex.org/publications/lexbib-a-corpus-and-bibliography-of-metalexicographical-publications/.

McCrae, J. P. et al. (2015): One ontology to bind them all: the META-SHARE OWL Ontology for the Interoperability of Linguistic Datasets on the Web. In: Gandon, F. et al. (eds.): The Semantic Web: ESWC 2015 Satellite Events, pp. 271–282. https://doi.org/10.1007/978-3-319-25639-9_42.

Wilkinson, M. D. et al. (2016): The FAIR Guiding Principles for scientific data management and stewardship. In: Scientific Data 3, p. 160018. http://doi.org/10.1038/sdata.2016.18.

Contact information

David Lindemann
UPV/EHU University of the Basque Country (Spain)
david.lindemann@ehu.eus

Penny Labropoulou
Institute for Language and Speech Processing (ILSP)/Athena R. C. (Greece)
penny@athenarc.gr

Christiane Klaes
TU Braunschweig (Germany)
University Library
c.klaes@tu-braunschweig.de

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Acknowledgements

The research presented in this article has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 731015 (ELEXIS, https://elex.is) and the COST Action NexusLinguarum - European network for Web-centered linguistic data science (CA18209), supported by COST (European Cooperation in Science and Technology, https://www.cost.eu/).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Christian-Emil Smith Ore/Oddrun Grønvik/Trond Minde

WORD BANKS, DICTIONARIES AND RESEARCH RESULTS BY THE ROADSIDE

Abstract Many European languages have undergone considerable changes in orthography over the last 150 years. This hampers the application of modern computer-based analysers to older text, and hence computer-based annotation and studies of text collections spanning a long period. As a step towards a functional analyser for Norwegian texts (Nynorsk standard) from the 19th century, funding was granted in 2020 for creating a full form generator for all inflected forms of headwords found in Ivar Aasen's dictionary published in 1873 (Aasen 1873) and his grammar from 1864 (Aasen 1864).

Creating this word bank led to new insight in Aasen (1873), its structure, internal organisation, and ambition level as well as its link to Aasen (1864). As a test, the full form list generated from this new word bank was used to analyse the word inventory of texts by Aa. O. Vinje, written in the period 1850–1870. The Vinje texts were also analysed using a full form list of modern standard Norwegian, to study the differences in applicability and see how Vinje's language relates to the written standard of modern Norwegian.

Keywords Dictionary and text analysis; full form systems; close reading of dictionaries

1. Introduction

Many modern European languages had their first written standard defined in the 19th century, but the standard may since have undergone substantial changes. This is the case for the Norwegian written standards. Until the second half of the 19th century, Danish was the only standard written language. The Nynorsk standard, based on the Norwegian vernacular, was introduced in the middle of the 19th century, but has been extensively revised. A history of standard revision causes modern tools for text analysis to be less well suited for older texts. Analysers must be adapted to both orthographic change and changes in inflectional morphology. This is a bootstrapping problem, since the creation of base forms with potential inflection forms requires analysis of text corpora, which in turn requires lemmatizers built from the same corpora.

Our solution to this problem for early Nynorsk is to use a central 19th century dictionary and grammar, both compiled by the Norwegian linguist and lexicographer Ivar Aasen (1813–1896). As a step towards a functional analyser for Norwegian texts (Nynorsk standard) from the 19th century, funding was granted in 2020 for creating a word bank and a full form generator for all headwords found in Ivar Aasen's Norsk Ordbog med dansk Forklaring 'Norwegian Dictionary with Danish definitions' (Aasen 1873) and Norsk Grammatik (Aasen 1864). The Aasen Word Bank was completed earlier this year.

Tools for making the full form generator were (1) Aasen (1864), (2) Aasen (1873), (3) the system and tool for Norwegian Word Bank (Norsk Ordbank), which is the main full form generator for modern Norwegian (separate ones for Bokmål and Nynorsk), see Hagen/Nøklestad (2010), Engh (2014), Grønvik/Ore (2014).

Creating the Aasen Word Bank led to new insight into Aasen (1873), its structure, organisation, and ambition level. The process and resulting findings will be discussed in this paper.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

2. The word bank system

The word bank is a structure for storing information about words and their inflected forms. The fundamental idea is that a word (lexical item) is identified as the set of all possible inflected forms. In this model, the headword used in a dictionary entry as base form is a representative for the set of word forms. The Norwegian written standards of Nynorsk and Bokmål have been revised several times after 1900, including changes in inflections. For example, in the Nynorsk standard "bok" (book) had until 2012 the additional form "bok" (det. sg. dem.). Since 2012, the form "boka" (det. sg. fem) is the only accepted form. Figure 1 shows a simplified word bank structure. The basic structure consists of (1) a list of base forms (headwords), (2) a table with the information about which paradigm(s) a base form can follow, and (3) a set of rewriting rules (paradigms). Table 2 is the pivot of the word bank. For a given base form, each corresponding row identifies a paradigm, the orthographic status of the corresponding forms and the timespan of this status. The paradigm table shows examples of the rewriting rules used to generate the inflected forms. The '+' is a wildcard character. The rewriting process runs as follows: The pattern in line 1 is used to find a match with a selected base form. In the example '+' will be bound to 'k'. Lines 2 to 4 are production lines, and the full form list will be "bok, boki, bøker, bøkene" and "bok, boka, bøker, bøkene". There is an overlap of forms. If required, the list of inflected forms can be reduced to unique forms with information about the corresponding line in the paradigm. All paradigms linked to a base form will have the same number of lines and marks on them.

Each paradigm has additional information about part of speech (POS), prototypical base forms and comments. In the Aasen Word Bank project every paradigm is referred to the relevant paragraph in Aasen (1864).

The word bank system was developed almost 30 years ago in connection with the construction of a rule based morpho-syntactic tagger for the two modern Norwegian written standards Bokmål and Nynorsk. The rewriting system is based on IBM's spellchecker, developed at the end of the 1980s. In the tagger project, base forms were linked to the corresponding entries in the general dictionaries Bokmålsordboka (Wangensteen 1986) and Nynorskordboka (Hovdenak et al. 1986). This was not required for the computational linguistic purpose of the taggers but has later turned out to be useful. The link between word bank entry and dictionary entry has made it possible to add a table of valid inflection forms to every entry in the online editions of these dictionaries. It is also possible to go from word bank entry to dictionary entry to check the definition(s) of a word, which is useful for separating homographs, etc.

Norwegian is a Germanic language with productive use of compounds, like for example German. A consequence is that the number of unique words (types) is for all practical purposes unlimited. A list of full forms can never be exhaustive. A solution to this problem is to use a so-called compound analyser, that is, a piece of software marking the border between possible elements. A compound analyser will be a future extension. We plan to test the software from the Oslo-Bergen tagger, see Hagen/Johannessen/Nøklestad (2000), Nøklestad (2022a, 2022b).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

(1) Base forms

Lemma	Base form
...	...
8701	bok
...	...

(2) Base forms, their paradigms and norm status

Lemma	Paradigm	In norm	From	To	...
...	...	...	...	...	...
8701	942	yes		2012.07.31	
8701	942	no	2012.07.31	31.12.9999	
8701	968	yes		31.12.9999	
...	...	...	...	...	...

(3) Rewriting rules for each paradigm

Paradigm	Line	Mark	Code
...	...	...	...
942	1	Sg indef	o+
942	2	Sg def	o+i
942	3	Pl indef	ø+er
942	4	Pl def	ø+er
...	...	...	...
968	1	Sg indef	o+
968	2	Sg def	o+a
968	3	Pl indef	ø+er
968	4	Pl def	ø+er
...	...	...	...

Fig. 1: Simplified structure of the word bank

The word bank structure has been used to create an inventory of Norwegian words with inflected forms and the history of changes in their orthographic status. The Word Bank is designed for modern Norwegian. It is a flexible structure, and by creating a new set of paradigm patterns based on Aasen (1864) and the headwords of Aasen (1873), it was possible to retro-create a word bank describing a language norm for the year 1873 – here termed the Aasen Word Bank.

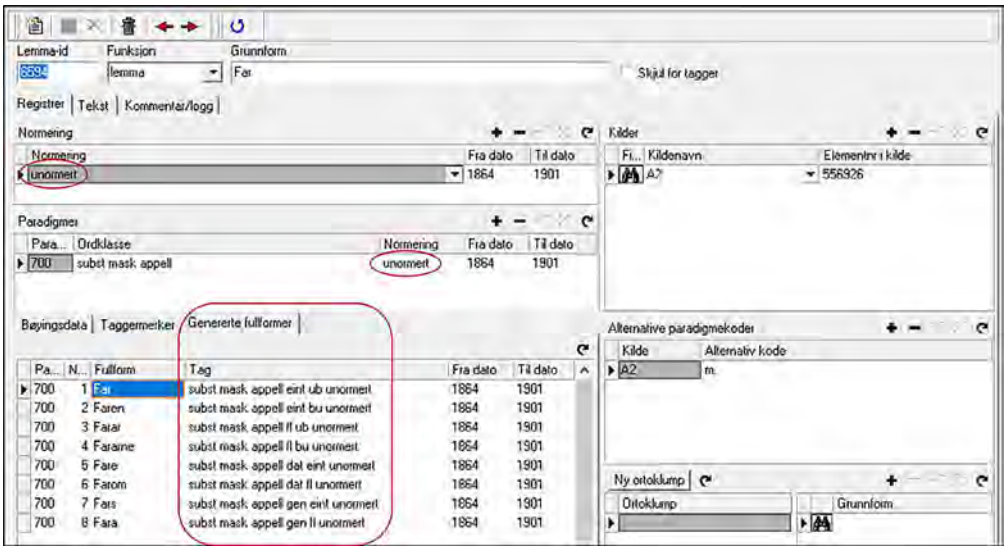


Fig. 2: Aasen Word bank: The headword "Far m" 'father'. Status "unormert" ('non-standard') for headword form and paradigm respectively in red frames

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

3. A description of Aasen's Norwegian Grammar (1864)

3.1 The contents and focus of Aasen (1864)

Aasen (1864) is a work of 394 pages with the contents ordered in introduction, five sections and two addenda. At the micro level, Aasen (1864) has 399 numbered paragraphs. Most paragraphs have a general section and a section for comments ("Anm."). Details pertaining to individual lexical items are found under "Anm.".

Aasen started reworking his "Grammatik for det Norske Folkesprog" (Aasen 1848) in the mid-1850s, but soon saw the need for substantial change. Aasen (1864) became an independent dissertation on the structure of Norwegian as a sum of spoken varieties. It introduces a standardised orthography and morphology based on an analysis of speech. Aasen (1864) aims at documenting the unity, coherence, and independence of the linguistic structure of Norwegian, as a Nordic language, related to but separate from Danish and Swedish; using Old Norse as a touchstone, including only what could be documented from Aasen's synchronic collections.

Aasen (1864) is rich in content, brief in style, and has little textual redundance. Every piece of information is given once; contents are logically ordered, resting on the assumption that the book has been read from the beginning, and that readers remember what has been read.

The macrostructure of Aasen (1864) is new, compared to Aasen (1848). Its five sections deal with (1) phonology, (2) base word form, (3) morphology, (4) word creation, (5) syntax.

Section 2 in the Grammar, Base word form, is new to Aasen (1864). This section discusses syllable structure and tone, includes a table of consonant clusters before and after the root vowel, and comments on the endings of disyllabic words for each part of speech (§ 75). Section 2 also deals with root vowel change, with a discussion of Ablaut and Umlaut in Norwegian. The systematic changes in base word forms from Old Norse to synchronic Norwegian and the relationship between Norwegian and its closest cognate languages is set out.

Much of the content of section 2 was inspired by a close study of Grimm's "Deutsche Grammatik" (1819–1837). In relation to Aasen (1873), section 2 sets out the framework for adapting headwords from Norwegian speech to model forms within a standard orthography, suggesting the categories of information to be expected in the treatment of individual words.

Section 4 (§ 241–300) discusses word creation in Norwegian, especially derivation from root forms or inflected forms, involving Umlaut, Ablaut, gender, and POS transitions. Aasen (1864) considers some frequently used word endings as compounding elements rather than derivational suffixes (§ 257–258). In the dictionary Aasen (1873), prefixes and suffixes regularly used in word formation have separate entries, including elements now used only in forming propria. The borderline between Aasen's view of derivation and compound can be explored further if the Aasen full form register is expanded with word forms from contemporary text, cf. chapter 5.

3.2 Aasen (1864) and the paradigm register

In the Aasen Word Bank each headword form is registered as it stands in Aasen (1873). Dialect forms (cross-referenced to other entries) that deviate from Aasen's standard orthography, are marked as non-standard headword forms, but may nevertheless have inflection

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

paradigms. This is possible because the entry format for the Word Bank allows separate marking for the headword form itself and its paradigm. This system also allows standard headword forms to be equipped with non-standard paradigms. For an example see Figure 2.

The most important section of Aasen (1864) in relation to the Aasen Word Bank is section 3" Bøiningsformer" ('Inflection Morphology'), comprising paragraph 151 to 240. Parts of speech (POS) are dealt with in sequence, starting with nouns and ending with verbs. Each POS has a general introduction; for nouns the issue of three linguistic genders comes first, then a discussion of linguistic gender versus biological gender. The categories of number, definiteness, and case in nouns are introduced as part of the inflection system for nouns. This system is used in relation to all word classes where inflection occurs. And since Norwegian is a Germanic language, nouns are also classified as "strong" (ending in a consonant) or weak (ending in a vowel).

Line nr	Marker	001 "kasta" 'throw'	309 "vera" 'be'
1	infinitive	a	vera
2	pres. sg.	ar	er
3	pres pl.	a	era
4	inf. passive	ast	verast
5	pret. sg.	ade	var
6	pret. pl.	ade	vaaro
7	supinum	at	voret
8	adj (perf. part.) neut. indef. sg.	at	voret
9	adj (perf. part.) neut. def. sg.	ade	vorna
10	adj (perf. part.) neut. indef. pl.	ade	vorne
11	adj (perf. part.) neut. def. pl.	ade	vorne
12	adj (perf. part.) masc. indef. sg.	ad	voren
13	adj (perf. part.) masc. def. sg.	ade	vorne
14	adj (perf. part.) masc. indef. pl.	ade	vorne
15	adj (perf. part.) masc. def. pl.	ade	vorne
16	adj (perf. part.) fem. indef. sg.	ad	vori
17	adj (perf. part.) fem. def. sg.	ade	vorna
18	adj (perf. part.) fem. indef. pl.	ade	vorna
19	adj (perf. part.) fem. def. pl.	ade	vorne
20	adj (pres. part.)	ande	verande
21	imperative sg.		ver
22	imperative pl.	e	vere
23	subjunctive sg.	e	vere
24	subjunctive pl.	e	vore

Fig. 3: Form registers for regular verbs of the type "kasta" 'throw', and the unicum "vera" 'be'

After the general POS introduction, individual inflection paradigms have a paragraph in which the recommended paradigm is shown in table form and commented on, in relation to (1) materials (speech forms), (2) Old Norse, and (3) Germanic cognates.

Aasen's systematic mapping of the inflection system of Norwegian made the registration of paradigms a comparatively easy task. When doubt arose about the shaping of a paradigm, an extra paradigm was created, rather than forcing an uncertain interpretation as the only possibility. Several base forms have more than one paradigm, each paradigm set with the status that seems appropriate.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Aasen (1864) starts with a maximum model for each word class and deals early with the most complex cases. The result is comprehensive and well-ordered paradigms for (frequent and well-documented) irregular nouns and verbs, but also a fair amount of over-classification, e. g. for members of regular verb groups, cf. the paradigm of verb paradigm 001 (Fig. 3, middle column) where many of the adjectival forms of the past participle are identical.

4. A description of Aasen's Norwegian Dictionary (1873)

Aasen (1873) is a bilingual dictionary of 964 pages with 38.742 entries and a couple of addenda. Headwords are in Norwegian, definitions and editorial language in Danish. It is designed to present Aasen's Norwegian vocabulary collections, covering the whole language and all dialects, in an acceptable form for a new Norwegian written standard, referred to as "Landsmaal"[1] 'the language of the country' or "Folkesprog" 'the language of the people'. Aasen's first grammar (1848) and dictionary (1850) of the Norwegian vernacular had served to establish Norwegian as a modern and independent language and himself as a noteworthy comparative philologist in the field of Nordic (Germanic) languages[2]. Aasen (1873) was his magnum opus and is regarded as the foundation of Norwegian lexicography and dialectology.

Aasen (1873) is also the companion volume to Aasen (1864). In the dictionary preface, Aasen discusses criteria for exclusion and inclusion of lexical items and word forms, underlining the need for brevity and clarity at all costs, but says nothing about the organization of the dictionary as a whole or within the entry. Aasen's biographers have discussed purpose, size, and orthography, but the only one to comment on lexicographical aspects is Dagfinn Worren (2006), and his comments concern alphabetical ordering and definition formats.

There is a good reason for this lack of interest. Detailed analysis of a dictionary presupposes that the contents can be classified and counted, a major task, and not one to be undertaken manually. The Aasen Word Bank facilitates analysis by numbers and reveals Aasen's working method in greater detail than what has been possible until now.

A nineteenth century dictionary is primarily a running text. To break up entries into ordered categories via a detailed lexicographic database would be forcing contents into a straitjacket and corrupting the result. It is better to respect the category system of the author, which in a dictionary means breaking the text into entries, isolating headwords with their POS, and then annotate entry contents. In this analysis, the aim has been to identify all word forms that Aasen himself presents as lexical items, whether in the standard orthography or in dialect form only.

4.1 Entry types, numbers, and contents

The entry formats of Aasen (1873) can roughly be divided into main (content) entries and cross-reference entries. The total number of lexical items (entry headwords and additional base forms with POS) extracted by machine analysis is 43.194. About 6000 additional head-

[1] The word form "Landsmaal" as a term was first used by Aasen in Aasen (1864 § 341).

[2] Aasen was well known to German philologists. The 1848 grammar and the 1850 dictionary were reviewed by Theodor Möbius in Gersdorfs Repertorium in February 1851, and both Grammar and Dictionary were sent to Möbius by Aasen himself immediately after publication. Aasen (1873) was reviewed by K. Maurer (1873) and F. Liebrecht (1874).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

words are so far identified by ongoing manual control, the total being likely to end at 50–53 000. This means that many entries cover more than one lexical item.

Each main entry gives a description of its headword according to Aasen's criteria. Inflected forms and a selection of speech variants are listed, as are forms from other Germanic languages and Old Norse. Aasen also illustrates the lexicogenetic potential of each headword by giving information about lexical items connected to the headword by form, i. e. derived forms, compounds and multiword expressions. A note on headword form as the first element in a compound is often included. Definitions are supported by synonyms or near-equivalents.

Some lexical items listed within entries of other headwords have their own entries, but not all, cf. Figure 4. Lexicographers of Norwegian after Aasen have tried to extract every lexical item in Aasen (1873), irrespective of the word form's placing as headword or within the entry text (see for example Grunnmanuskriptet (1935/1997)). The aim of the analysis in the Aasen Word Bank is the same. Doubtful cases are checked against the Norwegian language collections. A word form found as an independent lexical item, will be included. Doubtful cases particularly concern lexical items where the orthographic form may include spaces or hyphens.

(1) **Aar**, f. (Fl. **Aarar**), Aare, Redskab at roe med. G.N. ár (Eng. oar). Afvigende Fl. Aarer (Tel. og fl.). I Sammensætning sædvanlig med "a", som dog tildeels skulde være "ar"; saaledes: **Aara(r) blad**, n. Aareblad. **Aaraburd**, m. 1) den maade hvorpaa man bevæger Aarerne; 2) Rum til at røre Aarerne (Sfj.). **Aaradrag**, n. enkelt Drag eller Træk med Aarerne. **Aararlask**, s. Lask. **Aararlom** (oo), m. Grebet paa en Aare. **Aaraløysa**, f. Mangel paa Aarer.

Fig. 4: The entry for "Aar f." 'oar' lists six compounds, which can be seen as nested entries, or supplementary information to the headword Aar. The entry comments on the infix variation a/a(r)/ar. The Aasen Word Bank includes entries for compound prefixes plus infix.

4.2 Grouping headwords round a definition

In many cases Aasen's word harvest consists of several word forms from different parts of the country which are full synonyms, but not dialect variants of the same base form. The result may be an entry as shown in Figure 5, which is centred around a definition. Here, the structurally simplest form is selected as headword for the entry, while the other four are listed after the introduction "Ogsaa kaldet" 'also called'.

(2) **Kjøta**, f. Kjødside, Indside paa Skind eller Huder. Hard. Ogsaa kaldet: **Kjøkka** (Kjøtka), Hall., **Kjøtska**, Buskr. Ellers: **Kjøtrosa** (o'), f. B. Stift, Nordl. **Kjøtroslid** (i'), f. Sæt.

Fig. 5: The entry "Kjøta f." 'inside of (animal) hide'

Two of the other word forms are cross-referenced to this entry, the others are not mentioned anywhere else. In this type of entry, Aasen approaches the thesaurus entry format, which he knew from Roget's work. More than 600 entries include a synonym section similar to the entry above. This entry format came to attention through the Aasen Word Bank format, which makes it possible to sort out entries linked to multiple (standard and non-standard) base forms.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

4.3 Types of cross-reference entries in Aasen (1873)

The number of cross-reference entries in Aasen (1873) is about 4800 – 12,4% of the total number of entries (38.742). This is a very high share compared to later dictionaries describing a more standardized version of Nynorsk, where the proportion of cross-reference entries lies around 6% (the number is taken from the editorial database, dated 2013).

Most dictionaries indicate the status of a headword by the entry format. An entry consisting of a headword and a cross-reference to another headword, can be assumed to be less important, and most often a non-standard form of the target for the cross-reference entry. In Aasen (1873), this assumption would be a fallacy. The Aasen Word Bank has through its structure and comment system allowed a more detailed mapping of Aasen's use of the cross-reference system, carried out as part of the manual control. Cross-references are of two kinds, those which only point to a target headword (ca. 4100) and those which contain minimal other information about the headword, most often a Danish equivalent (ca. 700).

The simple type is used for linking dialect forms to standard forms (3), inflected forms to base forms (4), alternative standard forms to the main entry form (5):

(3) **abakleg**, s. avbakleg.

(4) **fraus**, s. frjosa.

(5) **andleg**, s. andeleg.

But often the headword form fails to match the target form linguistically, as in this example:

(6) **frestalle** (fleste), s. fleire. 'several'

There is no way the headword form "frestalle" can be a form of "fleire". The explanation is found under the entry "fleire":

(7) "Paa Sdm. **frestalle**, for flest-alle." 'In Sunnmøre **frestalle**, for "flest-alle'"

The expanded version would be "In one region of Norway, the multiword expression "flest alle" is used to express the notion 'several'". Cross-references in Aasen (1873) are also used as in encyclopaedias, meaning 'explanation or more information will be found under the target entry'.

A look in later dictionaries and the digital Norwegian language collections confirms that the standard forms "flest alle" and "flestalle" are listed as a separate lexical item. In the Aasen Word Bank an entry for "flest-alle" has been added with links to the two entries "frestalle" and "fleire". This means that all cross-reference entries must be checked manually, to ensure that each headword is correctly marked in the Aasen Word Bank as standard or non-standard.

Cross-references can also serve to draw attention to the most widely attested speech form. For some lexical items, Aasen for system reasons chose a disyllabic headword form, although the dominant spoken form is monosyllabic. In the dictionary (Aasen 1873) – the opening of the entry for "Fader" as standard form:

(8) **Fader**, m. (Fl. **Feder**), Fader (pater). Lyder mest alm. **Faer**, **Far** (som i Svensk og Dansk); … .

The most widely used speech form "Far" is cross-referenced to "Fader m" (and another entirely different headword).

(9) **Far**, m. s. Fader og Fare.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

The simple form "Far" then turns up in compounds for relationship words, with uncertain status information. Orthographic reformers after Aasen who preferred a more orthophone approach, would therefore often find their work done for them in Aasen (1873), with information on extent and usage of the non-standard forms.

4.4 To what extent is Aasen (1873) a normative dictionary?

The headwords of main entries in Aasen (1873), taken together, are generally considered a proposal for a Nynorsk standard orthography. Although Nynorsk had been used in books and journals since the 1850s, its literary corpus was still very limited in 1873, and the orthography was heterogenous. Aasen knew very well that he was putting forward a proposal, not a decree. He could hope for acceptance, but not command it. Aasen's authenticity principle dictated that he would not include a word unless he himself had heard it or had information on form, sense and usage confirmed from more than one contemporary source whom he trusted. His language collections – now in the National Library – start in the 1830s and gain force in the 1840s. In addition to his own collections, he received several manuscripts from others. From these, he used what he could verify, if he trusted the informant.

In the introduction to Aasen (1873), Aasen states that although there is some literature written in forms of modern Norwegian, it is not included in the materials used for the dictionary (p. XII). His dictionary is a presentation of spoken, contemporary Norwegian, properly verified.

Aasen's textual corpus was transcribed speech, in dialect form. Modern Norwegian yet had no written standard expression. The model he used for synthetizing dialect forms into a proposed standard orthography was to find an orthographic form which would link speech forms in a systematic way. The consonant digraph "rn" as syllable ending is the standard example. In Aasen's day, "rn" had been replaced in speech by /dn/ (sound differentiation) or /nn/ (assimilation). But "rn" would explain both speech forms, had the support of use in standard Danish and Swedish, and was used in Old Norse, so "rn" was introduced, e. g. in "Bjørn" 'bear'.

Aasen set out to document the whole language. This means that his collections are solid and multi-sourced for frequent vocabulary, but also rich in thinly documented and doubtful cases. His approach to identifying dialect forms with a standard form, and putting the standard form into its proper linguistic context, is revealed by studying entry types and entry format.

4.5 Aasen (1873) – dictionary type and mode of analysis

In his discussion of bilingual dictionary types, the lexicographer Ladislav Zgusta outlines three subgroups "with a remarkably outstanding concentration upon some purpose" (Zgusta 1971, pp. 304 f.). One of them he terms the "ethnolinguistic bilingual dictionary" constructed for languages with little or no written literature. In such cases, the defining language is often another, well-established literary language, used to bridge a culture gap. A dictionary of this type is designed to introduce as a written standard a language existing only in the vernacular and might well be termed a pioneer dictionary.

Aasen (1873) is a pioneer dictionary, designed to introduce a written standard for the Norwegian vernacular (Grønvik 1992). Several of the dictionary features discussed above be-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

come methodically rational in this perspective. His materials are heterogeneous in terms of form and place of origin – the dictionary must convince its users that the language is cohesive and well designed. The well-documented central vocabulary is the backbone of the dictionary. Words that are not so well attested, must be guided into the language system through cross-referencing and by being included as support material in more comprehensive entries. Cross-referencing irregular inflection forms guides users to the proper entries. By organising his dictionary in this fashion, Aasen is also able to throw light on the word creation system of Norwegian, especially by establishing the prefix form of compounds.

The least conventional feature, seen from present-day lexicographical practice, is the type of entry that groups linguistically unrelated synonyms round a definition, without giving the headwords in the group conventional entries, or even cross-references. Aasen (1873) is a semasiological dictionary, but this type of entry belongs in an onomasiological dictionary. Aasen was deeply interested in onomasiology and left a manuscript for a Norwegian thesaurus published posthumously (Norsk Maalbunad 1925).

The Aasen Word Bank was launched because it is needed in analysing early Nynorsk text. But analysing the dictionary through the Word Bank system has also brought new insights in the dictionary as it stands, in Aasen's lexicographical method (which he never discussed in any context), and above all made the contents of the dictionary more accessible and therefore possible to evaluate. It is to be hoped that the Aasen Word Bank will contribute to a wider scholarly interest in and use of Aasen (1873).

5. A first application of the Aasen Word Bank as an analytic tool

The Nynorsk written standard of today is based on Norsk Grammatik (1864) and Norsk Ordbog (1873), two major works by Ivar Aasen. In 1885, the Norwegian Parliament accepted Nynorsk as a second written standard in addition to Danish.

The works of the Norwegian author Aa. O. Vinje (1818–1870) are early examples of texts written in a form of Nynorsk, with a degree of standardisation. Aasen (1873) was published three years after the death of Aa. O. Vinje, so Vinje cannot have used it. But Vinje and Aasen were in close contact for many years. It is generally assumed that Aasen advised Vinje on his orthography (Aasen 1957), although final choices will have been Vinje's own. The most important choices on inflection morphology are published in a small supplement to his journal Dølen (Vinje 1859).

We have used text written in Norwegian from the collected works of Vinje (Vinje 1916–1921) as a first test case for the form list produced form the Aasen word bank. Vinje's primary literary output is journalism and essays. In the mid-19th century, the tradition of long novels was not developed in Norway, in contrast to what we find in France and UK. The collection of Vinje's works is not large, only 750.000 running words.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

	Word forms found	Unique word forms found
In both	51.8977	10.968
Only in Aasen Word Bank	35.371	4.572
Only in the Nynorsk Word Bank	44.409	7.837
Total found in either and/or both	598.757	23.377
Not found in neither	79.312	24.493

Table 1: The resulting numbers from the analysis of Vinje's writings

From the Aasen Word Bank a list of all full forms was generated, approximately 290.000 unique word forms or 520.000 unique triples (word form + POS + information about inflection categories). A similar list with recommended forms generated from the word bank of modern Nynorsk consists of 415.723 unique word forms and 585.046 triples (word form + POS + information about inflection categories). The Aasen Word Bank has 45.000 base forms, and the modern Word Bank has 120.000. As we have seen, the Aasen norm has a much richer inflectional system, which explains the relatively larger number of word forms in the Aasen word bank.

A small program identifies the word forms in the running text and checks for each if it is an inflected form in the Aasen Word Bank and/or an inflected form of modern Norwegian. Table 1 shows the resulting numbers. As mentioned in section 2 the word bank does not have a system for compound analysis. Consequently, the number of matches is lower than it could be.

The total number of unique word forms is 47.870. Of these, 26.917 occur once, while 157 word forms occur 500 times or more. A brief comment on the most frequent word forms: 137 of them are found in both word banks, nine only in the Aasen Word Bank, three only in the Nynorsk Word Bank, and seven word forms are specific to Vinje. Of the nine in the Aasen Word Bank only, six are inflection forms where the spelling has been changed after 1873, the last three base forms. The three in the Nynorsk Word Bank only are inflection forms changed after 1873. Some of the word forms specific to Vinje can today be seen as signature forms ("ikki" 'not' (instead of Aasen's "ikkje"), "ero" 'are' (plural form used in Aasen 1864, but replaced with the form "era" in Aasen 1873)). Others reflect his linguistic environment, and the occurrence of Danish text in his writings (the Danish form "ikke" 'not', the preposition form "af" which he used in Norwegian and Danish irrespectively). Others just seem to show different orthographic habits from what Aasen was to recommend in his dictionary three years after Vinje's death.[3]

About half, 49%, is found in one or both of the word banks used. 51% is not found in either word bank. A finer classification has been done on the word forms of the letter a (1798 instances) Below we comment briefly on word forms starting with a- and found 1) in both word banks, 2) only in the Aasen Word Bank, 3) only in the Nynorsk Word Bank, 4) in neither word bank.

1) The unique word forms from Vinje that are found in both word banks comprise the core vocabulary of Nynorsk, with inflected forms. The number of compounds is relatively

[3] Vinje wrote "kver" 'every', Aasen "kvar"; "altid" vs. "alltid"; "up" vs. "upp".

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

small. Imported vocabulary is not there, because Aasen did not include it in his dictionary. There are very few doubtful cases, mainly because Aasen's orthographic norm is very clearly defined and made explicit by the Aasen Word Bank.

2) The word forms from Vinje found only in the Aasen Word Bank are mainly inflected forms that are no longer part of Nynorsk orthography, e. g. regular adjectives ending in "-ad" (now reduced to "-a").

3) Of the word forms from Vinje found only in the Nynorsk Word Bank, about 40% are forms of imported (e. g. non-Germanic) vocabulary (not included in Aasen's dictionary, but essential in journalistic text), 17% name forms, ca. 10% then classified as Danish word forms, since included in Nynorsk standard orthography. The rest, about 33%, are word forms of headwords that are either consistent with Aasen's orthography, but missing in his dictionary (25%), or specific to Vinje's personal and much more heterogeneous orthography (8%). Vinje wrote, printed, and sold twice a week, and when in doubt he seems to have chosen word forms consistent with his dialect from Western Telemark, some of which have found their place in modern Nynorsk.

4) Of the word forms found in neither word bank, a much larger sample should have been analysed than there has been time for, for the word creation system for Norwegian gives great weight to frequently used particles. Vinje often used the preposition word form "af" where Aasen chose "av" 'of'; this choice affects ca. 300 compounds. It can however be said with certainty that the larger groups of word forms specific to Vinje will belong to the following types: a) Danish word forms (from quotes, and shorter texts in Danish). b) Imported word forms, many of which today have an orthography adapted to Norwegian. c) Word forms consistent with Aasen's orthography, but not found in Aasen's dictionary. d) Word forms specific to Vinje, i. e. word forms spelt by the ear and often reflecting Vinje's dialect basis.

There is also a sprinkling of (quoted) word forms from modern languages and Latin. Vinje was proud of his command of English, and this group seems to be the largest.

The conclusion must be that Vinje and Aasen mostly agreed on what their written (standard) Norwegian should look like. In the frequently used word forms, the number of deviations between Vinje and Aasen are few, but they show up because of their frequency. As for the rest, Vinje was a writer and journalist navigating in uncharted waters, his concern was to make sure he got read. He used and made the words he needed, and when in doubt, it seems that Norwegian speech was his compass.

References

Aasen, I. (1848): Det norske Folkesprogs Grammatik. Kristiania.

Aasen, I. (1850): Ordbog over det norske Folkesprog. Kristiania.

Aasen, I. (1864): Norsk Grammatik. Kristiania.

Aasen, I. (1873): Norsk Ordbog. Med dansk Forklaring. Kristiania.

Aasen, I. (1925): Norsk maalbunad. Samanstilling av norske ord etter umgrip og tyding. Oslo.

Aasen, I. (1957): Brev og dagbøker. B. 1. Brev 1828–1861. [ed. Reidar Djupedal]. Oslo.

Engh, J. (2014): IBMs leksikografiske prosjekt for norsk 1984–1991. In: Maal og Minne 106 (1): pp. 67–101. http://ojs.novus.no/index.php/MOM/article/view/225. (last access: 25-03-2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Grimm, J. (1819–1837): Deutsche Grammatik. Göttingen

Grunnmanuskriptet (1935/1997): https://usd.uib.no/perl/search/search.cgi?tabid=993&appid=59.

Grønvik, O. (1992): The earliest dictionaries of Nynorsk in the light of present day dictionary typology. In: The Nordic Languages and Modern Linguistics 7. Proceedings of the Seventh International Conference of Nordic and General Linguistics in Tórshavn, 7–11 August 1989, Vol. I. Føroya Fróðskaparfelag (Annales Societatis Scientiarum Færoensis, Supplementum XVIII). Tórshavn.

Grønvik, O. (2016): The lexicography of Norwegian. International handbook of modern lexis and lexicography. Berlin/Heidelberg, pp. 1–34.

Hagen, K./Johannessen, J. B./Nøklestad, A. (2000): A constraint-based tagger for Norwegian. In: 17th Scandinavian Conference of Linguistics, Volume I, no. 19. Odense Working Papers in Language and Communication.

Hagen, K./Nøklestad, A. (2010): Bruk av et norsk leksikon til tagging og andre språkteknologiske formål. LexicoNordica (17). https://tidsskrift.dk/lexn/article/view/18624 (last access: 25-03-2022).

Hovdenak, M. et al. (1986): Nynorskordboka. definisjons- og rettskrivingsordbok. [3. ed. 2006]. Oslo.

Nøklestad, A. (2022): The Oslo Bergen Tagger. https://github.com/noklesta/The-Oslo-Bergen-Tagger (last access: 25-05-2022).

Nøklestad, A. (2022): The Compound analyzer software. https://github.com/textlab/mtag (last access: 25-05-2022).

Ore, C.-E. S. (2016): Gamle ordbøker og digitale utgaver. Nordiske Studier i Leksikografi 13. Rapport fra 13. Konference om Leksikografi i Norden København 19.–22. mai 2015. København, pp. 203–216.

Ore, C.-E. S. (2020): Å ta Hans Ross på ordet: Ross' ordbok i relasjon til Aasens med Metaordboka som verktøy. Nordiska studier i lexikografi 15.Rapport från 15 konferensen om lexikografi i Norden. Helsinki, pp. 253–264.

Roget, P. M. (1852): Thesaurus of English words and phrases classified so as to facilitate the expression of ideas and to assist in literary composition. London.

Venås, K. (1996): Då tida var fullkomen. Ivar Aasen. Oslo.

Vinje, Aa. O. (1859): Det norske Landsmaals vigtigaste Bøygningsformer. Bilag til Dølen nr. 30, 1859.

Vinje, Aa. O. (1916–1921): Skrifter i Samling I–V. Oslo.

Walton, S. J. (1996): Ivar Aasens kropp. Oslo.

Wangensteen, B. (1986): Bokmålsordboka. Definisjons- og rettskrivningsordbok. [3. ed. 2005]. Oslo.

Worren, D. (2006): Molbech som mønster for Aasen. In: Nordiske Studier i Leksikografi 8. Nordisk Forening for Leksikografi, pp. 391–406.

Contact information

Christian-Emil Smith Ore
Department of Linguistics and Scandinavia Studies, University of Oslo
c.e.s.ore@iln.uio.no

Oddrun Grønvik
Department of Linguistic, Literary and Aesthetic Studies, University of Bergen
Oddrun.Gronvik@uib.no

Trond Minde
Department of Linguistic, Literary and Aesthetic Studies, University of Bergen
Trond.Minde@uib.no

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Ana Ostroški Anić/Ivana Brač

AirFrame
Mapping the field of aviation through semantic frames

Abstract The paper presents the process of developing the AirFrame database, a specialized lexical resource in which aviation terminology is defined in the form of semantic frames, following the methodology of the Berkeley FrameNet (FN). First, the structure of the database is presented, and then the methodology applied in developing and populating the database is described. The link between specialized aviation frames and general language semantic frames, of which frames defining entities, processes, attributes and events are particularly relevant, is discussed on the example of the semantic frame of Flight and its related frames. The paper ends with discussing possibilities of using AirFrame as a model for further developing resources in which general and specialized knowledge are linked.

Keywords Terminology; aviation terminology; semantic frames; specialized knowledge; specialized lexicography

1. Introduction

Aviation is a professional domain in which language plays a crucial role in ensuring regular daily operations and safe communication. There is hardly any other professional environment in which the communicative setting is as described and prescribed as the one of the international aviation community. However, aviation language consists not only of aviation terminology and radiotelephony phraseology in English, but it also includes a certain level of general language vocabulary that is necessary for effective professional communication. Defining the lexical component of aviation language is therefore a challenging task because of all the different areas of expertise encompassed, but also because not all aviation subdomains are equally relevant for effective communication (Bratanić/Ostroški Anić 2010). In a similar manner, it could be claimed that any domain of specialized knowledge is characterized by different types of categories that are often intertwined with categories of general knowledge and human experience used in a specialized context.

The need for linking general and specialized knowledge has been well addressed in developing specialized resources based on the theory of Frame Semantics (Faber et al. 2011; L'Homme/Subirats/Robichaud 2016; L'Homme/Robichaud/Subirats 2020; Pilitsidou/Giouli 2020) or on its terminological application in the form of Frame-Based Terminology (Faber et al. 2011; Faber/Buendía Castro 2014; Faber 2015). Unlike traditionally organized specialized dictionaries or terminological databases that define professional knowledge through hierarchically organized data categories, interpreting specialized categories as dynamic structures calls for providing a more dynamic approach to the processing and presentation of terminology (Faber 2015). Although semantic relations between frames and their elements are also largely subject to a hierarchical structure, they nevertheless allow for including different chronological and associative relations (Ruppenhofer et al. 2016).

AirFrame is a specialized lexical resource in which the domain of aviation is defined in terms of semantic frames, i. e. specialized aviation frames are linked to general language semantic frames, of which top level frames defining various entities, processes, attributes

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

and events are most relevant (Brač/Ostroški Anić 2019). The AirFrame database was designed with a view of defining a specialized domain, but at the same time linking it to more general linguistic information, paving the way for concept modelling that enables linking resources of a different origin and a different purpose. This paper presents the structure of the database and the methodology applied in developing it, with a particular focus on links between general and specialized semantic frames defined in it. After the introduction, a brief overview of the application of FrameNet's methodology to defining specialized knowledge is given. The structure of semantic frames and their elements is presented in the third section, while the methodology of frames identification and description is given in the fourth section of the paper. Discussion is based on the example of the sematic frame of Flight and its related frames. The paper ends with suggestions for possible applications of the database data.

2. Application of FrameNet to specialized domains

There have been many applications of Frame Semantics to lexicography since Fillmore first defined the *frame* as "a system of categories structured in accordance with some motivating context" (Fillmore 1982), identifying experience as the context necessary for successful categorization. In the fields of terminology, specialized lexicography and specialized knowedge representation, a number of applications of Frame Semantics to specialized resources have been developed in order to enable a more accurate linguistic and computational representation of the conceptual level of specialized knowledge.

Kiktionary (Schmidt 2007) is an English-German-French database of football vocabulary extracted from the corpus of reports from football matches, in which FrameNet's methodology is combined with the Wordnet's principle of organizing synonyms and homonyms in synsets. BioFrameNet (Dolby/Ellsworth/Scheffczyk 2006) is another database applying the FrameNet's methodology in processing texts in the field of biomedical sciences and molecular biology. The domain of law is represented in, among others, the Italian resource combining Frame Semantics and Van Kraling's approach to the representation of law (Venturi et al. 2009) and JuriDiCo (Pimentel 2015), a database of Portuguese and English legal terminology modelled in accordance with the methodologies applied in FrameNet and the Canadian database DiCoInfo (L'Homme 2012).

Frame-Based Terminology emerged as a theoretical approach in terminology studies, applying the principles of Frame Semantics to terminology work by developing a model of dynamic description of categories of specialized knowledge, (Faber Benítez/Márquez Linares/Vega Expósito 2005; Faber 2015) where focus is put on the description of a prototypical event in a specialized domain, e. g. the environmental event in the domain of environment (Faber/Buendía Castro 2014). The prototypical event serves as a general frame for the organization of more specific concepts, and it encompasses macro-categories as "concept roles characteristic of this specialized domain" (Araúz/Reimerink/Faber 2009). In Puertoterm, a knowledge base on environmental engineering, an event includes a natural agent and a human agent, a process with the subcategories of a natural and an artificial process and a construct, and a patient/result. The human agent can use an instrument to create an artificial process or constructs (Araúz/Reimerink/Faber 2009). In Ecolexicon, an environmental knowledge base, the same dynamic conceptualization of the event structure is applied on all levels, which means that terminographic definitions can be also considered as "mini-knowledge representations or frames" (Araúz/Reimerink/Faber 2009, p. 51).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Termframe, a terminology knowledge base for the field of karstology, is another resource developed on the principles of Frame-Based Terminology, containing terms and their definitions in English, Slovene and Croatian. In Termframe, a definition template is defined for each concept category in the domain model (Vintar/Stepišnik 2021), following the frame-based approach that views a definition as a small frame of knowledge. What makes Termframe particulary relevant for our work is the development of first frame-based definition templates for Croatian, which could be adapted for terminology work in another specialized domain.

The work done by L'Homme and colleagues in developing domain-specific resources as applications of the Frame Semantics methodology and theoretical principles (L'Homme/Robichaud/Rüggeberg 2014; L' Homme/Robichaud/Subirats 2020) was of greatest influence in devising the AirFrame's methodology and data categories structure. Their model of linking general language semantic frames with domain-specific frames relies on using a set of 15 semantic roles, general enough in their description to be applicable to a number of terms in more semantic frames (Pimentel/L'Homme/Laneville 2012), and it served as the starting point for developing a methodology of linking aviation related semantic frames to their general top-level instances. The next section describes the structure of the AirFrame database in more detail.

3. The structure of AirFrame

AirFrame is the first specialized frame-based lexical resource in the Croatian language, consisting of aviation related semantic frames and frame elements (FEs) with their accompanying definitions and examples, types of frame elements, lexical units and frame-to-frame relations.[1] Since the domain of aviation is typically characterized by numerous events, activities and well-controlled processes, FrameNet's methodology of describing knowledge categories in terms of hierarchically related events and activities they structure seemed particularly appropriate for the identification of this field.

In designing the data categories structure, we tried to adhere to the FrameNet's structure as close as possible, but at the same time to use the categories that reflect the basic principles of terminology work. Apart from administrative categories, each frame therefore consists of a frame definition, frame elements with their definitions, examples of sentences in which the frame specific frame elements are used, lexical units and relations to other frames. AirFrame is primarily a database of Croatian aviation terminology, so the definitions of frames and frame elements are written in Croatian, as are the types of frame elements and frame-to-frame relations. However, examples of frame elements and all lexical units are given for Croatian, English and French alike.

Following the FrameNet distinction, frame elements can be core, which are frame specific and inessential in the realization of a given frame, and non-core, i. e. elements that do not uniquely characterize the frame (Ruppenhofer et al. 2016), but further define it by placing it in a certain time and space. A further distinction between non-core and peripheral elements is not applied, nor are extra-thematic elements defined.

Another divergence from the FrameNet model is the omission of annotated examples for FEs lexical realization. Although the annotation of sentences as examples of lexical units in

[1] AirFrame is available at airframe.jezik.hr.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

linguistic context has been done, it is not yet integrated into the database. The category of frame element example therefore does not illustrate valency patterns for particular lexical units that can appear in syntactic position of a defined element, as it does in FrameNet, but it resembles the category of context as a typical element of a traditional terminology database. A defined FE is still placed within an actual linguistic context, but the difference between a terminological context and an example of an FE in AirFrame lies in the fact that the example can contain any of the lexical units – i. e. all synonyms and variants of a term – of the same meaning in the frame. All instances of examples are corpus examples, and can be given for Croatian, English and French. There is a difference in the approach to example illustration between aviation specific frames and general language frames, which is discussed in the next section.

Since no ontology has yet been developed as a formal model of conceptual representation in AirFrame, the category of the frame element type was introduced to serve as the connection between the semantic and ontological levels, i. e. as an implicit top level ontology. A set of 17 frame element types was composed on the basis of FrameNet's semantic types, EuroWordNet's top level entities (Vossen 2002) and the semantic roles defined in the LIRICS project (Petukhova/Bunt 2008). Unlike the ontological semantic types for the frame elements in FrameNet that categorize the sort of filler expected in the element (Ruppenhofer et al. 2016, p. 86), the FE type attributed to an FE in AirFrame classifies the kind of role the FE is. FE types in that sense serve as the superordinate macro roles for the frame specific semantic roles or frame elements, and are used to group elements in conceptually connected groups. E. g., the frame `Flight` has several frame elements referring to space and time: AIRSPACE and AERODROME are core elements bearing the FE type of location, as well as the non-core element FLIGHT_ROUTE.[2] The core element FLIGHT_TIME and non-core elements FLIGHT_DURATION, FREQUENCY and TIME_SPAN are marked with the FE type time.

Semantic frames are invoked by lexical units, which in AirFrame can be words of general language or aviation terms, and all instantiate FEs in appropriate frames. They are entered into the database separately from the frames, so they can be attributed to more than one frame. Lexical units are defined for Croatian, English and French, but only grammatical information for each is entered. When compared to the terminological information in a traditional termbase, lexical units would correspond to terms, while frame elements would roughly be on the level of concepts. However, a one-to-one relation cannot be established since FEs have the role of frame specific semantic roles.

Finally, all frames are linked by 13 FrameNet's frame-to-frame relations. Some relations are more relevant for the identification of conceptual relations between domain categories, such as *has subframe* and *subframe_of*, corresponding to partitive terminological relations, as well as *inherits* and *is_inherited_by*, which could be compared to generic relations, e. g. the type_of relation. Other relevant frame relations are shown in section 5.

4. Methodology for defining specialized semantic frames in AirFrame

The process of describing aviation semantic frames can be broadly divided into two large segments: identifying frames and defining them. Before starting any terminology work,

[2] The names of frames are written in fixed-width font, `Courier New`, while the names of frame elements are written in SMALL CAPS.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

regardless of the type of resource envisaged as the final product, one needs to first have a general overview of the domain whose terminology is being processed. The development of AirFrame was made easier in that sense because it continued from the work previously done on the Croatian aviation terminology (Ostroški Anić 2020).

The field of aviation was first broadly divided into large categories according to the basic processes and entities included and connected to the central event of the domain, i. e. the flight. Most aviation training material, e. g. handbooks, manuals and guidance material, are organized in a similar fashion, i. e. in a way which follows an aircraft from pre-flight to post-flight activities. This top-down approach is supported by a corpus-based analysis, for which two corpora had been compiled.

First, a parallel English-Croatian corpus was compiled using documents from the Directory of legal acts of the European Union, chapter Transport policy, subchapter Air transport in English and Croatian. Out of 220 documents from the Air transport subchapter, 178 legal acts were taken having both (English and Croatian) language versions. The texts were downloaded from the EUR-Lex database, and entered in the Sketch Engine's corpus compilation module (Kilgarriff et al. 2014). The English part of the corpus consists of a little over 950 000 words, while the Croatian part consists of 855 000 words. A monolingual corpus of aviation related texts in Croatian was compiled, too, for which the available textbooks, manuals, reports, scientific papers and dissertations, as well as student diploma papers and MA theses in Croatian were used. The aviation corpus in Croatian consists of 2,210,000 words.[3]

The parallel corpus was first used for term extraction and validation of term candidates' lists (Ostroški Anić/Lončar/Pavić 2019). An automatic term extraction was conducted for each language with the option of extracting a list of 1000 single-word and multi-word keywords. The EUR-Lex English 2/2016 corpus was used as a reference corpus for extracting the English single-word term candidates, while the English Web 2013 was used as a reference corpus for extracting the multi-word term candidates. Similar options were possible for term extraction in Croatian. Manual analysis and term verification of both English and Croatian candidate lists of extracted terms was then conducted.

After acquiring a list of terms or lexical units from the parallel corpus, a similar process was done for the Croatian aviation corpus. The list of terms extracted from the monolingual corpus served for the validation of terms from the parallel corpus. Given the nature of the legal discourse of the Eur-Lex documents in the parallel corpus, few definitions of aviation concepts were used in the description of semantic frames and their elements. The Croatian aviation corpus was used for this instead.

The semantic frames were then identified by grouping the extracted lexical units according to the aviation concepts they denote. In FrameNet, the basic criterion for delimiting one frame from another is that all lexical units should "evoke the same type of event and share the same inventory and configuration of FEs" (Ruppenhofer/Boas/Baker 2014). Therefore, lexical units of the same semantic type, and appearing as arguments of the same verbs should be placed in the same frame. Determining the scope of the semantic frame, however, largely depends on the granularity of the conceptual description, which is, on the other hand, conditioned by the potential use of the database.

Adding general language frames to AirFrame is done for two reasons: the first is the top-level categorization that provides a continuity in relations between more general frames and

[3] Both corpora are available to Sketch Engine users by contacting the authors.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

specialized frames as their instantiations. The second reason is laying down the foundations for a future FrameNet of the Croatian language, which could use the existing top-level frames, and enrich them if necessary.

The frames of general language are of the same structure as the aviation frames, except that they are not identified in the same way, but are taken over from FrameNet and adapted for Croatian. If all the frame elements of a certain frame from the Berkeley FN are appropriate for the description of the Croatian counterpart, nothing is changed. If the Croatian syntactic description asks for an additional element or a change in the existing FEs structure, changes are introduced accordingly. The Croatian general language corpus hrWaC is used for the examples of FEs (Ljubešić/Klubička 2016). English and French examples of FEs and lexical units, otherwise added in aviation semantic frames, are not entered in the general language frames. Users can look for this information in the English and French FrameNet.

5. The semantic frame of Flight

Flight is without doubt the central event in the domain of aviation. There are several perspectives one can take in order to define its activities, processes and entities engaged. Flight can be defined as an instance of travel, in which case we are taking the view of passengers and defining it as the period of transport by aircraft from boarding the airplane to its disembarkation. Flight can also be defined by taking into account the cargo and baggage being transported, but in AirFrame it is described with regard to the use of the aircraft by authorized aviation personnel to conduct the activity of operating the aircraft for flying from one location to another.

As already said, every frame has a definition that contains all core FEs that are conceptually necessary to understand the frame. The definition of `Flight` in AirFrame is: The AIRCRAFT moves from a specific AERODROME to a specific destination AERODROME through AIRSPACE for a specified DURATION. The definition tells us that the core frame elements are: AIRSPACE, AERODROME, AIRCRAFT, and DURATION, while non-core elements include, among others: PILOT, ALTITUDE, FLIGHT_ROUTE, FLIGHT_PATH, FLIGHT_CONDITIONS, FLIGHT_CREW, MANNER, FLIGHT_SPEED, FREQUENCY, TIME_SPAN, etc. FEs are inferred from corpus examples, and are defined in relation to a specific frame, e. g. AIRCRAFT is defined as 'a device that can maintain istelf and move in atmosphere', FLIGHT_SPEED as 'the speed of the aircraft in relation to the Earth's surface', etc. Below every definition of a frame element, there are examples in Croatian, English, and French to show its use in context, as shown in Figure 1.

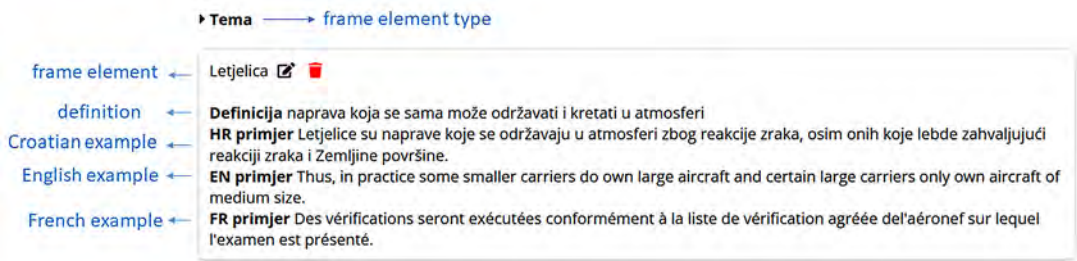


Fig. 1: Frame element *Letjelica* (eng. Aircraft)

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Different frame elements are grouped under frame element types, understood here as macro roles connecting semantic and ontological levels of information. In `Flight`, the frame element type Location encompasses AIRSPACE and AERODROME as core elements of `Flight`, and FLIGHT_ROUTE as a non-core element. MANNER and SPEED fall under the FE type Manner, and FREQUENCY and TIME_SPAN under Time. A general FE type of Theme is attributed to the core-element AIRCRAFT when used in that semantic role, while the PILOT is a specification of the FE type of Agent.

Although semantic annotation of corpus examples showing the argument structure of FEs is going to be added to the database in the next phase of the development, a significant number of sentences has already been annotated in a separate database. Examples (1) to (6) show the annotation of sentences with the target units FLIGHT and FLY. Frame elements are marked with a subscript, while frame element types are written in superscript.

(1) Avioni predviđeni za let na visinama iznad 25,000 stopa moraju biti opskrbljeni jedinicama za raspodjelu kisika.

$^{\text{THEME}}[_{\text{AIRCRAFT}}$ Airplanes] intended for FLIGHT $^{\text{LOCATION}}[_{\text{ALTITUDE}}$ at altitudes above 25,000 feet] must be equipped with oxygen distribution units.

(2) Ako pilot zrakoplova radi sigurnosnih ili hitnih letačko-operativnih razloga smatra da nastavak leta prema izvornom aerodromu odredišta nije preporučljiv, može preusmjeriti let na drugi aerodrom, kojega smatra prikladnim.

If $^{\text{AGENT}}[_{\text{PILOT}}$ the pilot of the aircraft], $^{\text{CAUSE}}[_{\text{CAUSE}}$ for safety or emergency flight-operational reasons], considers that the continuation of the FLIGHT $^{\text{FINAL_LOCATION}}[_{\text{AERODROME}}$ to the original destination aerodrome] is not recommended, $^{\text{AGENT}}[_{\text{PILOT}}$ he] may redirect the FLIGHT $^{\text{FINAL_LOCATION}}[_{\text{AERODROME}}$ to another aerodrome] which he considers appropriate.

(3) Putanja dopušta da helikopter nastavi let od visine krstarenja do visine od 300 m (1000 ft) iznad helidroma.

$^{\text{THEME}}[_{\text{FLIGHT_PATH}}$ The trajectory] allows $^{\text{THEME}}[_{\text{AIRCRAFT}}$ the helicopter] to continue its FLIGHT $^{\text{INITIAL_LOCATION}}[_{\text{ALTITUDE}}$ from the height of the cruise] $^{\text{FINAL_LOCATION}}[_{\text{ALTITUDE}}$ to an altitude of 300 m (1000 ft) above the heliport].

(4) Finnair je putnicima koji su u SAD letjeli preko Helsinkija jedno vrijeme poklanjao SharpWizard 8000.

$^{\text{AGENT}}[_{\text{AIRLINE}}$ Finnair] gave away ^{THEME}[SharpWizard 8000] ^{GOAL} $[_{\text{PASSENGER}}$ to passengers] who FLEW $^{\text{FINAL_LOCATION}}$[to the USA] $^{\text{LOCATION}}[_{\text{FLIGHT_ROUTE}}$ via Helsinki] $^{\text{TIME}}[_{\text{TIME}}$ for a while].[4]

(5) Kandidat mora letjeti helikopterom najmanje 5 sati noću, od čega najmanje 3 sata s instruktorom, uključujući 1 sat rutnog navigacijskog letenja te 5 samostalnih polijetanja i slijetanja sa zaustavljanjem.

$^{\text{AGENT}}[_{\text{PILOT}}$ The candidate] must FLY $^{\text{THEME}}[_{\text{AIRCRAFT}}$ the helicopter] $^{\text{TIME}}[_{\text{DURATION}}$ for at least 5 hours at night], of which $^{\text{TIME}}[_{\text{DURATION}}$ at least 3 hours] $^{\text{AGENT}}[_{\text{PILOT}}$ with an instructor], including $^{\text{AMOUNT}}[_{\text{DURATION}}$ 1 hour] $[_{\text{FLIGHT}}$ of route navigational flight] and $^{\text{AMOUNT}}[_{\text{AMOUNT}}$ 5] $[_{\text{FLIGHT_SEGMENT}}$ solo take-offs and landings with stopping].

(6) Prvi američki mlazni zrakoplov, Boeing B-707, proizveden je 1958. godine i letio je uglavnom preko Sjevernog Atlantika.

$^{\text{THEME}}[_{\text{AIRCRAFT}}$ The first American jet, the Boeing B-707], was built $^{\text{TIME}}[_{\text{TIME}}$ in 1958] and FLEW mostly $^{\text{LOCATION}}[_{\text{FLIGHT_ROUTE}}$ across the North Atlantic].

[4] The frame element type Goal also covers the semantic roles Recipient and Beneficiary, which are attributed to animate participants of an event.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

There are many lexical units evoking the frame `Flight`, as can be seen in Figure 2. There is no terminological preference given to certain lexical units over others, the way terms are usually classified according to normative preference in a traditional terminological resource. All synonyms and variants of a given term can therefore be added as lexical units because they all do evoke a semantic frame for which the term is relevant. However, spelling and formal variants are usually omitted.

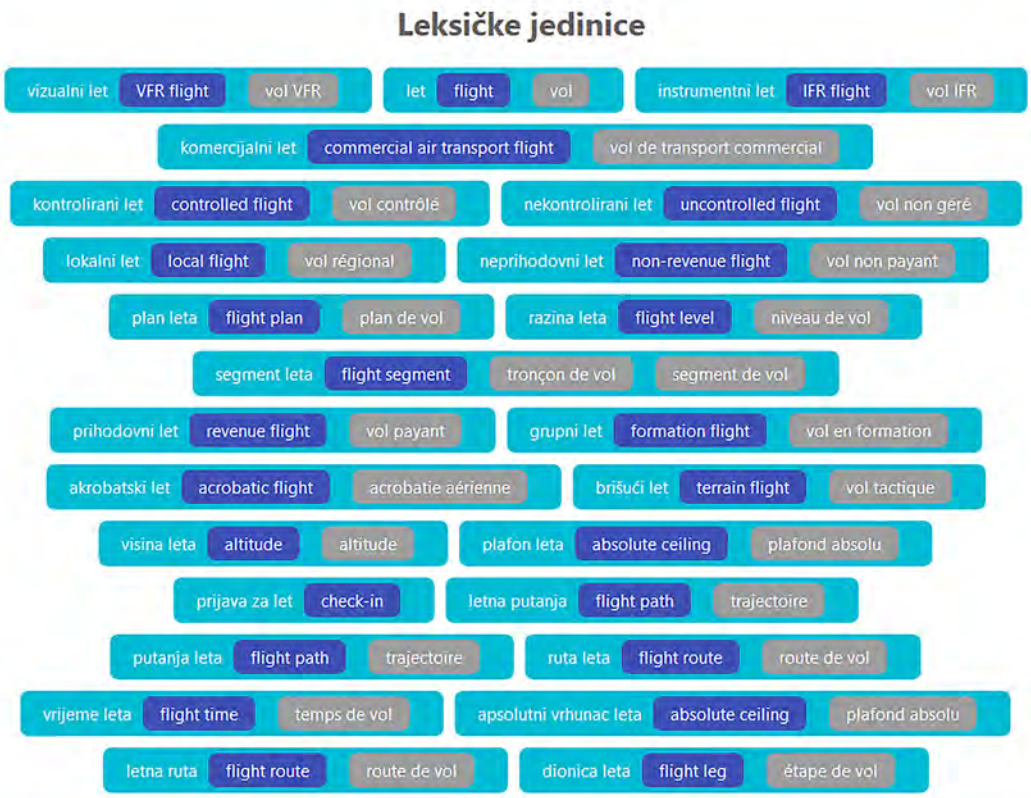


Fig. 2: Croatian, English and French lexical units of the frame `Let` (eng. `Flight`) represented in the search engine airframe.jezik.hr

Since one of the aims of AirFrame is to present the structure of the field of aviation as comprehensively as possible, frame-to-frame relations have been included in the frame description. `Flight` inherits all FEs from top-level frames `Event` and `Motion` as its parent frames, with more specific FEs added. Being a complex frame that an event is, `Flight` can be further divided into subframes, i.e. into six subframes that correlate to the phases of flight: `Take-off`, `Landing`, `Climb`, `Cruising`, `Descent`, `Approach`. Finally, `Flight` uses certain elements of the frames `Airspace` and `Airport`, while the frame `Flight_travel` is closely related to it and therefore linked by the relation *see_also*. Frame-to-frame relations used to connect `Flight` to related frames in the AirFrame database are shown in Figure 3.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Fig. 3: Frames related to `Flight`

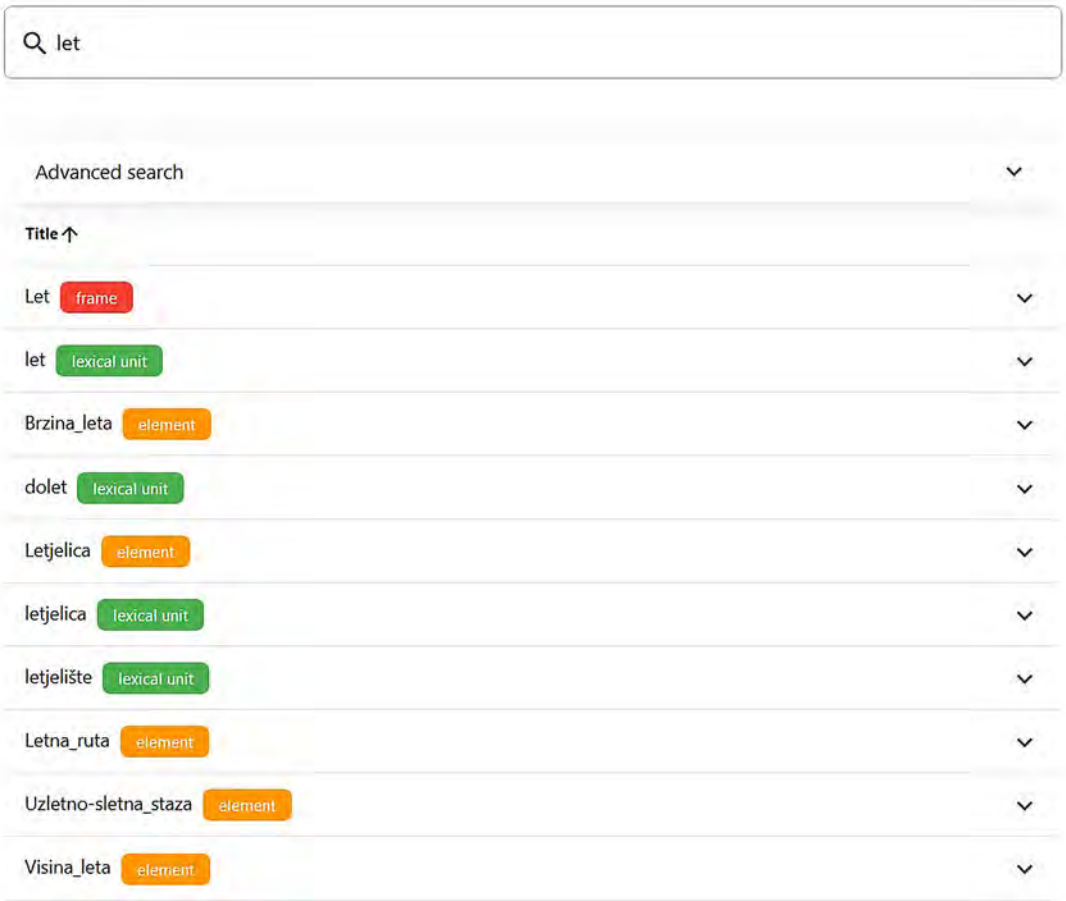


Fig. 4: Search results for *let* 'flight' in the AirFrame search engine

The analysis of such a complex category that combines different spatial and temporal relations is a good example of the possibility of connecting extra-linguistic and linguistic

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

knowledge in the form of semantic frames, which can be used in a number of different applications. When used as a teaching and reference resource, AirFrame's offers a more outlined presentation of the conceptual level of specialized knowledge than users find in traditional specialized resources. As can be seen in Figure 4, users can search for any term related to the frames, frame elements and lexical units defined, while the display of results visually keeps the distinction between different categories in the database, as well as between lexical units in different languages.

6. Concluding remarks

Specialized knowledge defined and presented in a domain-specific lexical resource such as the one presented here provides ample opportunities for further use. However, the envisaged users of AirFrame, as well as users of all similar specialized resources, are not typical terminology users relying on intuitive data categories presentation that enables them to quickly find answers to various terminological issues. Domain lexical resources are often more used by machines than humans, perhaps because they provide a thorough representation of a complex network of semantic relations, which lies underneath all linguistic structures. Their potential is nevertheless more and more recognized in developing LSP courses and teaching material. Although students of aeronautics, aviation engineering, air traffic and related studies, as well as translators working in these fields, are without doubt primary intended users of AirFrame, the database is being developed as a model of specialized knowledge description that can be used in the description of other, related specialized domains.

AirFrame follows the footsteps of several similar specialized resources that apply Frame Semantics and versions of the FrameNet's methodology in particular. What sets it apart is the introduction of the level of general semantic roles that can be used for linking different resources that use semantic role labelling, whether those are general language or domain-specific resources. Using the frame specific semantic roles gives us a detailed insight into the structure of categories, and allows for mutual relations to be established between different semantic frames.

Grouping frame specific elements into higher level semantic roles, on the other hand, gives us a list of all elements of the same type that appear in different syntactic functions. For example, the lexical units *airplane*, *helicopter* or *jet*, which in the examples presented in the previous section are defined as belonging to the FE AIRCRAFT, and to which the FE type Theme is assigned, may appear in the subject's position in cases where no agent is expressed. In other words, it is not expressed that someone flies the aircraft. In sentences where an agent is expressed, i. e. in sentences where the FE PILOT is realized by the lexical units *pilot* and *candidate*, the agent appears in the object position. Based on the comparison of all FE within the same FE type, we can thus make conclusions on different syntactic patterns in which they appear, and whether a change in the lexical unit used for a particular FE changes the intended meaning of the whole utterance.

A combination of a fine-grained semantic description with an enabled level of generalization might pave the road to the development of a robust conceptual model that preserves terminological information in the integration of terminological resources into larger linguistic networks.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

References

Araúz, P. L./Reimerink, A./Faber, P. (2009): Puertoterm & Marcocosta: a frame-based knowledge base for the environmental domain. In: Proceedings of the XVIII FIT World Congress. Shanghai, p. 24.

Brač, I./Ostroški Anić, A. (2019): From concept definitions to semantic role labeling in specialized knowledge resources. In: Gürlek, M./Naim Çiçekler, A./Taşdemir, Y. (eds.): Proceedings of the 13th International Conference of the Asian Association for Lexicography, pp. 604–611.

Bratanić, M./Ostroški Anić, A. (2010): Compiling lexical information for an aviation English dictionary. In: Reconceptualizing LSP. Online proceedings of the XVII European LSP Symposium 2009.

Dolby, A./Ellsworth, M./Scheffczyk, J. (2006): BioFrameNet: a domain-specific FrameNet extension with links to biomedical ontologies. In: Bodenreider, O. (ed.): Proceedings of KR-MED. Formal Biomedical Knowledge Representation. Baltimore, pp. 87–94.

Faber Benítez, P./Márquez Linares, C./Vega Expósito, M. (2005): Framing terminology: A process-oriented approach. In: Meta: Journal des traducteurs 50 (4). https://www.erudit.org/fr/revues/meta/2005-v50-n4-meta1024/019916ar/ (last access: 29-06-2022).

Faber, P. (2015): Frames as a framework for terminology. In: Kockaert, H. J./Steurs, F. (eds.): Handbook of terminology. Amsterdam, pp. 14–33. http://doi.org/10.1075/hot.1.02fra1.

Faber, P. et al. (2011): Linking specialized knowledge and general knowledge in EcoLexicon. In: Actes de la conférence Terminologie & Ontologie: Théories et Applications (TOTh) 2011. Annecy, pp. 47–61.

Faber, P./Buendía Castro, M. (2014): EcoLexicon. In: Abel, A./Vettori, C./Ralli, N. (eds.): Proceedings of the 16th EURALEX International Congress. Bolzano, pp. 601–607. https://euralex.org/publications/ecolexicon/.

Fillmore, C. J. (1982): Frame semantics. In: Linguistics in the morning calm. Selected papers from SICOL-1981. Seoul, Korea, pp. 111–137.

Kilgarriff, A. et al. (2014): The Sketch Engine: ten years on. In: Lexicography 1 (1), pp. 7–36. http://doi.org/10.1007/s40607-014-0009-9.

L 'Homme, M.-C./Robichaud, B./Subirats, C. (2020): Building multilingual specialized resources based on FrameNet: application to the field of the environment. In: Torrent, T. T./Baker, C. F./Czulo, O./Ohara, K./Petruck M. R. L. (eds.): International FrameNet Workshop 2020 Towards a Global, Multilingual FrameNet. Marseille, pp. 85–92.

L' Homme, M.-C./Subirats, C./Robichaud, B. (2016): A proposal for combining "general" and specialized frames. In: Proceedings of the 5th Workshop on Cognitive Aspects of the Lexicon (CogALex – V). Osaka, Japan, pp. 156–165. https://aclanthology.org/W16-5321.

L'Homme, M.-C. (2012): Adding syntactico-semantic information to specialized dictionaries: an application of the FrameNet methodology. In: Lexicographica 28, pp. 233–252. http://doi.org/10.1515/lexi.2012-0012.

L'Homme, M.-C./Robichaud, B./Rüggeberg, C. S. (2014): Discovering frames in specialized domains. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14). Reykjavik, Iceland, pp. 1364–1371. http://www.lrec-conf.org/proceedings/lrec2014/pdf/455_Paper.pdf.

Ljubešić, N./Klubička, F. (2016): Croatian web corpus hrWaC 2.1. In: http://nlp.ffzg.hr/resources/corpora/hrwac/, https://www.clarin.si/repository/xmlui/handle/11356/1064.

Ostroški Anić, A. (2020): Hrvatsko zrakoplovno nazivlje od leksikografije do terminologije. In: Brač, I./Ostroški Anić, A. (eds.): Svijet od riječi. Terminološki i leksikografski ogledi. Zagreb, pp. 361–374.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Ostroški Anić, A./Lončar, M./Pavić, M. (2019): Extracting lexical units for identifying specialized semantic frames. In: Terminologija 26, pp. 73–87.

Petukhova, V./Bunt, H. (2008): LIRICS semantic role annotation: design and evaluation of a set of data categories. In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08). Marrakesh, pp. 39–45.

Pilitsidou, V./Giouli, V. (2020): Frame semantics in the specialized domain of finance: building a termbase to translation. In: Gavriilidou, Z./Mitsiaki, M./Fliatouras, A. (eds.): Euralex XIX. Lexicography for inclusion, pp. 263–271. https://euralex.org/publications/frame-semantics-in-the-specialized-domain-of-finance-building-a-termbase-to-translation/.

Pimentel, J. (2015): Using frame semantics to build a bilingual lexical resource on legal terminology. In: Kockaert, H. J./Steurs, F. (eds.): Handbook of Terminology. Amsterdam, pp. 427–450. http://doi.org/10.1075/hot.1.usi1.

Pimentel, J./L'Homme, M.-C./Laneville, M.-È. (2012): General and specialized lexical resources: a study on the potential of combining efforts to enrich formal lexicons. In: International Journal of Lexicography 25 (2), pp. 152–190. http://doi.org/10.1093/ijl/ecr025.

Ruppenhofer, J. et al. (2016): FrameNet II: extended theory and practice. https://framenet2.icsi.berkeley.edu/docs/r1.7/book.pdf (last access: 07-03-2022).

Ruppenhofer, J./Boas, H. C./Baker, C. (2014): The FrameNet approach to relating syntax and semantics. In: Gouws, R. H./Heid, U./Schweickard, W./Wiegand, H. E. (eds.): Dictionaries. An international encyclopedia of lexicography. Berlin/New York, pp. 1320–1329.

Schmidt, T. (2007): The Kicktionary: a multilingual lexical resource of the language of football. In: Data Structures for Linguistic Resources and Applications: Proceedings of the Biennial GLDV Conference 2007. Tübingen, pp. 189–196.

Venturi, G. et al. (2009): Towards a FrameNet resource for the legal domain. In: Proceedings of the 3rd Workshop on Legal Ontologies and Artificial Intelligence Techniques: 2nd Workshop on Semantic Processing of Legal Text., p. 10. http://ceur-ws.org/Vol-465/paper8.pdf.

Vintar, Š./Stepišnik, U. (2021): TermFrame: a systematic approach to Karst Terminology. In: Dela (54), pp. 149–167. http://doi.org/10.4312/dela.54.149-167.

Vossen, P. (2002): EuroWordNet General Document.
https://research.vu.nl/ws/portalfiles/portal/77020259/EWNGeneral.

Contact information

Ana Ostroški Anić
Institute of Croatian Language and Linguistics
aostrosk@ihjj.hr

Ivana Brač
Institute of Croatian Language and Linguistics
ibrac@ihjj.hr

Acknowledgements

This work has been fully supported by the Croatian Science Foundation under the project UIP-2017-05-7169.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Kristel Proost/Arne Zeschel/Frank Michaelis/
Jan Oliver Rüdiger

MAP (Musterbank Argumentmarkierender Präpositionen)

A patternbank of argument-marking prepositions in German

Abstract Recent years have seen a growing interest in linguistic phenomena that challenge the received division of labour between lexicon and grammar, and hence often fall through the cracks of traditional dictionaries and grammars. Such phenomena call for novel, pattern-based types of linguistic reference works (see various papers in Herbst 2019). The present paper introduces one such resource: MAP ("Musterbank argumentmarkierender Präpositionen"), a web-based corpus-linguistic patternbank of prepositional argument structure constructions in German. The paper gives an overview of the design and functionality of the MAP-prototype currently developed at the Leibniz-Institute for the German Language in Mannheim. We give a brief account of the data and our analytic workflow, illustrate the descriptions that make up the resource and sketch available options for querying it for specific lexical, semantic and structural properties of the data.

Keywords Argument structure; valency; prepositions; constructicography; construction grammar

1. Argument structure

Learning which participants of a given scene must be expressed in a sentence, and how they must be coded formally, is an essential part of learning any language. Within the framework of Valency Theory, argument structure (or: valency) is understood as a lexical property of particular words (notably verbs). Correspondingly, it has usually been covered in specialised dictionaries which list such properties for individual predicate heads (cf. e.g. VALBU 2004; VDE 2004). However, recent years have seen growing objections to a purely word-centred ('lexicalist') approach to argument structure, since it fails to capture relevant semantic generalisations across the syntactic patterns of different verbs. This has led to a revised view of argument structure patterns such as [SUBJ V OBJ-1 OBJ-2] as meaningful linguistic signs in their own right ('constructions' in the Construction Grammar sense, cf. Goldberg 1995, 2006). In lexicography and grammaticography, this idea has inspired a range of approaches that seek to complement the traditional formats of valency lexicography with construction-based descriptions ('constructicography', cf. Lyngfelt et al. 2018). Such descriptions are not only of interest for the scientific community in linguistics, but also hold great promise for practical applications in language pedagogy, where they can serve to spell out 'hidden' aspects of grammatical meaning for second language learners.

For means of illustration, in the German valency dictionary VALBU, the verb 'search, look for' is said to select for a PP headed by the preposition *nach* 'after' that refers to an object which the searching agent strives to obtain. However, the same is true for many other verbs that are not included in VALBU, but also exhibit a 'strive to obtain'-reading when combining with a *nach*-PP. For instance, *Duden Online* gives the following paraphrases for *nach etwas tauchen, nach etwas bohren* and *nach etwas scharren*, respectively: 'tauchend nach etwas

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

suchen, etwas zu erreichen, zu finden suchen'[1] ('to look for something, try to attain or find something by diving'), 'mithilfe von Bohrgeräten nach etwas suchen'[2] ('to look for something with the help of drilling equipment') and 'scharrend nach etwas suchen, durch Scharren aus der Erde o. Ä. zu fördern suchen'[3] ('to look for something by scratching, to try to extract something from the earth etc. by scratching'). When such descriptions are considered side by side, it becomes obvious that the shared meaning is more plausibly ascribed to the shared argument structure pattern [V + *nach* + NP] ~ 'attempt to attain NP by V-ing' (presumably inherited from the prototype *suchen nach* NP) than to three special (and each time independent) verb readings of *tauchen* 'dive', *bohren* 'drill' and *scharren* 'scratch'.

In order to represent such generalisations in a suitable kind of linguistic reference work, novel formats of description are required that are no longer organised around an inventory of words (predicate heads), but around an index of the complex constructions in which these words occur (argument structure patterns).

2. The MAP patternbank

MAP is a new patternbank of German that seeks to provide such descriptions for argument structure patterns with a prepositionally marked argument. We focus on descriptions for constructions with the ten most common prepositions with local (i. e., spatial) source meaning (*an, auf, aus, bei, in, nach, über, um, von, vor*). In this section, we describe our workflow in devising these descriptions (2.1) and give a brief overview of the web-based resource in which they are provided (2.2).

2.1 Data and workflow

Our descriptions are based on data from the German reference corpus DeReKo, which contains 53bn tokens of (mostly newspaper) texts from Germany, Austria and Switzerland. For each preposition, we start out by setting up a preliminary semantic classification of its argument structure patterns based on a literature survey and the data from a large prepositional valency dictionary (see section 2.1.1). Next, a random sample of 1m sentences per preposition is drawn and cleaned in a semi-automatic pre-processing procedure (2.1.2). Third, a sample of 10,000 instances is drawn from the remaining data and annotated for a variety of lexical, semantic and grammatical properties. In the course of this process, the initial semantic classification scheme is usually substantially revised (2.1.3). When data coding is finished, we produce the actual descriptions that are contained in the resource, often resulting in further adjustments to pattern boundaries in order to secure a comparable descriptive grain size.

[1] "tauchen" in Duden online. https://www.duden.de/node/180296/revision/419356 (last access: 04-05-2022).

[2] "bohren" in Duden online. https://www.duden.de/node/132381/revision/542512 (last access: 04-05-2022).

[3] "scharren" in Duden online. https://www.duden.de/node/162037/revision/472136 (last access: 04-05-2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

2.1.1 Preliminary semantic classification

As a first step, we compare existing descriptions of a given target preposition in specialised preposition dictionaries (Kiss et al. 2016; Schmitz 1964; Schröder 1986) and grammars that deal with such constructions in some detail (Helbig/Buscha 2017; Schulz/Griesbach 1982; Weinrich 2005; Zifonun/Hoffmann/Strecker 1997). In addition, we extract all verbal valency patterns from an electronic version of Müller's (2013) valency dictionary of German prepositions and devise a preliminary semantic classification of these patterns.[4]

Classifications are formulated as paraphrases involving up to three pattern arguments and the pattern-specific semantic relation that holds between these arguments. The patternbank only covers constructions in which the PP does not have a purely local or temporal adjunct meaning, since such uses of the prepositions in question are already well-documented in many dictionaries and grammars. By contrast, grammaticalised uses as a verbal object marker have long been viewed as essentially meaningless in German linguistics (Duden: Die Grammatik 2016, p. 618; Engelen 1975, p. 111; Pittner 1999, p. 50). Although this view is no longer generally accepted, systematic and sufficiently detailed characterisations of the semantic relations that hold between the arguments of such patterns are as yet wanting (though see Höllein 2019 for a recent research monograph that takes some steps towards closing this gap).

Pattern arguments are labeled with the perspectival roles FIGURE (i. e., trajector, located element) and GROUND (i. e. landmark, reference point). Causative variants of such prepositional FIGURE-GROUND relations include an additional EFFECTOR:

(1) a. *Die Firmen*$_{\text{FIGURE}}$ *bohren nach Öl*$_{\text{GROUND}}$.
'The companies drill for oil'

b. *Die Regierung*$_{\text{EFFECTOR}}$ *lässt die Firmen*$_{\text{FIGURE}}$ *nach Öl*$_{\text{GROUND}}$ *bohren.*
'The government makes the companies drill for oil'
Pattern meaning: '<FIGURE> attempts to obtain <GROUND>'

In this manner, all pattern meanings are represented in a unified format consisting of two or three arguments (FIGURE and GROUND, sometimes augmented with an additional EFFECTOR) and the pattern-specific semantic RELATION (invoked by the predicate expression, if only indirectly sometimes) that holds between these elements (here: 'attempt to obtain'). Before we identify these patterns in our own data, some further preparatory steps are taken.

2.1.2 Preprocessing

From the initial sample of 1m sentences per preposition, a number of instances are removed ahead of coding. These include:

– Hits in which the putative preposition is actually a verb particle or product of a tokenisation error

[4] In the case of *nach*, Müller lists more than 500 patterns with verbal heads. Pattern descriptions consist of an entry (e. g., "<jemand taucht nach etwas (Dat.)>" 'somebody dives for something (dative)'), a meaning paraphrase ('jemand begibt sich unter Wasser auf der Suche nach dem Genannten', 'somebody goes underwater in search of something') and one or more associated corpus examples (e. g. "Damon Edmunds taucht im Haigebiet vor Südaustralien nach Meeresmuscheln" 'Damon Edmunds is diving for seashells in the Southern Australian shark area'; Müller 2013, p. 1957). Prepositions that are more frequent (such as *in* 'in') may have more than three times as many patterns than *nach* in Müller's dictionary.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

- Hits in verbless sentence tokens
- Hits in which the preposition forms part of a set of pre-identified multiword expressions that are irrelevant for our purposes (as devised from idiom dictionaries and other phraseological resources as well as n-gram-analyses, e. g. *nach wie vor* 'still')
- Hits in which the preposition is part of a name (e. g. place name, *Frankfurt am Main*)
- Hits in which the preposition heads a nominal attribute[5] (e. g. *Sehnsucht nach* 'longing for')
- Hits in which the preposition heads the complement of an adjective (e. g. *süchtig nach* 'hooked on')
- Hits that are putatively purely local in meaning, since the preposition combines with a motion or posture verb (e. g. *nach Frankfurt fahren* 'drive to Frankfurt', *nach links neigen* 'lean to the left', cf. section 2.1.1)
- Hits in which the preposition heads a temporal (or other clear) adjunct (e. g. *nach 10 Uhr losfahren* 'leave after 10 o'clock, cf. section 2.1.1)

With quite some variability between the different prepositions, these cases usually add up to at least 25% of the original data. These false positives are then removed in order to relieve subsequent coding. From the remaining data, a random sample of 10,000 instances is drawn on which the descriptions in the patternbank are based.

2.1.3 Coding

These 10,000 samples are then coded manually by at least two annotators for the following properties:

- Pattern meaning
- Predicate properties: lexical head, predicate type (verbal, complex, verbal + reflexive, complex + reflexive), predicate diathesis (active, passive, converse), realised verb complements (other than the three pattern arguments), occurrence in a construction with special argument realisation properties (e. g., infinitival constructions, imperatives etc.)
- FIGURE properties: grammatical function of the FIGURE argument (e. g. "subject"), syntactic category of the FIGURE argument (e. g. "subject clause")
- GROUND properties: prepositional head (e. g. "nach"), governed case of the complement (e. g. "dative"), lexical head of the complement (e .g. "Öl" in an expression like *nach* Öl *bohren* 'to drill for oil')
- EFFECTOR properties (where relevant): grammatical function of the EFFECTOR argument (e. g. "subject"), syntactic category of the EFFECTOR argument (e. g. "subject clause")

During coding, preliminary semantic classifications are adjusted as appropriate in order to accommodate new, hitherto non-anticipated usages. Also, it is common for expressions to be grouped and regrouped several times over as new semantic commonalities come to the attention of the annotators in the course of the classification task.

2.2 The resource

The contents of the patternbank are both textual (comprising different types of articles for descriptions on different levels of abstraction, see below) and diagrammatic (comprising

[5] At a later stage, we plan to include these attributive usages in the patternbank, too.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

visualisations of different kinds of formal and semantic usage preferences, cf. section 2.2.1). A complex faceted search allows pattern instances to be filtered for a wide variety of semantic, lexical and structural features (cf. 2.2.2).

2.2.1 Contents

Prepositional argument structure patterns can be described on different levels of abstraction. The meaning associated with the verb-preposition combination *nach etwas bohren* 'to strive to obtain something by drilling' in (1), for example, is also evoked by combinations of the same preposition with other verbs: *nach etwas kratzen/tauchen* equally evoke the meaning 'strive to obtain', where the verbs specify the manner in which the FIGURE strives to obtain the GROUND (by scratching or diving, respectively). Slightly similar meanings arise from combinations of *nach* with *schlagen* 'to hit', *werfen* 'to throw' and *treten* 'to kick' and from syntagms consisting of *nach* and *sich sehnen* 'to yearn', *lechzen* 'to crave' and *dürsten* 'to thirst'. While the former express the meaning 'strive to reach someone/something' (cf. Levin 1993; Perek 2015), the latter evoke the meaning 'strive to experience'. The three meanings could in principle be taken to call for three different patterns, or for only one abstract pattern whose meaning could be paraphrased as 'to strive for something'.

The patternbank provides descriptions on different levels of semantic abstraction in order to deal with grain size issues of this kind. *Low-level descriptions* ("pattern articles") constitute the core of the patternbank and provide an account of individual prepositional argument structure patterns with their specific meaning, attested formal realisations in the German reference corpus DeReKo and typical usage patterns in the data. *Mid-level descriptions* ("family articles") capture semantically coherent families of such patterns that can be differentiated from other families with the same preposition but different meanings. *Top-level descriptions* ("marker articles") systematise the full spectrum of families with their associated member patterns, list individual patterns not belonging to any overarching family and address the relation of the abstract pattern-specific meanings of the preposition to its local source meaning. These are the most abstract/inclusive descriptions contained in the resource.

As indicated, the patternbank also contains visualisations of quantitative usage preferences of the relevant prepositional argument marker in the investigated sample (10,000 attestations). These include the relative frequencies of different argument structure patterns and pattern families that share the same prepositional argument marker, the relative frequencies of specific syntactic realisations of a given pattern, the relative frequencies of different case realisations of the GROUND argument (where variable) and the relative frequencies of different lexical fillers of the elements RELATION and GROUND.

2.2.2 Query options

A powerful faceted search permits users to identify samples that match a broad variety of semantic, lexical and structural properties. *Semantic properties* can be filtered with a tag system for categories such as 'causal', 'scalar', 'adversative' etc. *Lexical properties* include the lemmatised head of the sample's predicate expression, its predicate type (reflexive like *sich erkundigen* 'to inquire' vs. non-reflexive such as *ringen* 'to wrestle', simple like *fragen* 'to ask' vs. complex such as *Frage stellen* 'to ask a question') as well as the internal constituents of complex predicate expressions (e. g. V+NP.acc for verbonominal predicates such as *Frage stellen* 'ask a question', *Feuer fangen* 'catch fire' etc.). *Structural properties* include

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

various aspects of the formal make-up and grammatical marking of the four pattern elements RELATION, EFFECTOR, FIGURE and GROUND. These include the syntactic function of the EFFECTOR and FIGURE arguments, the governed case of the GROUND argument and the diathesis type of the RELATION element (= predicate expression, i. e. 'active', 'passive' or 'converse'). Apart from the internal structure of predicate expressions and the grammatical function of their arguments, it is possible to filter for a range of constructions that may alter the canonical argument realisation of a given pattern (such as e. g. various non-finite constructions and subjectless imperatives).

The query options also allow for semantic, lexical and structural properties to be combined such that users may e. g. search for instances where predicates of a given type are used in a specific syntactic construction and/or with a particular diathesis type and/or a GROUND element in the accusative or dative case. Likewise, users may search for instances of patterns with particular semantic properties selecting a tag like 'causal' or 'scalar' and then filter instances of these patterns for realisations of the EFFECTOR or FIGURE arguments in a particular syntactic function.

In the following section, we give an overview of the structure and contents of pattern-level descriptions ("pattern articles"), which form the backbone of the resource.

3. Pattern articles

Pattern articles contain information on the meaning and form of prepositional argument structure patterns as well as their instantiating predicates. These three types of specifications are described in separate sections of the articles.

3.1 Meaning section

The meaning section gives a paraphrase of the pattern-defining relation between FIGURE and GROUND (i. e., the overall scenario designated by the pattern, e. g. 'strive to obtain') and the semantic type of the arguments. If the canonical realisation of the pattern involves an EFFECTOR in addition to the two core arguments FIGURE and GROUND, this is also stated and exemplified in the meaning section.

In case the pattern shows semantic variation, the meaning paraphrase generalises over all meaning variants. Meaning variants are listed in a separate subsection. Variants typically arise from differences in the semantic type of the pattern arguments (entity vs. event or state of affairs, concrete entity vs. abstract entity etc.).

Pattern instances that are ambiguous between two or more meaning variants (*Er fragte nach seinem Gehalt* 'He asked for his salary' vs. 'He asked about his salary') are listed and commented on in a separate paragraph of the meaning section. The relation of the pattern-specific abstract meaning of the preposition to its local source meaning is also addressed in a separate subsection ('conceptualisation').

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

3.2 Form section

All formal realisations of a given pattern in the underlying sample (10.000 sentences) are listed in the form section of the pattern articles.[6] Formal realisations are first grouped according to the case of the GROUND element (insofar as there is variability here) and subsequently for their diathesis type (e.g. first all variants with governed accusative case, and within these, first all active variants, then all converse ones, then all passive ones etc.). Within these subsets of expressions, formal realisations are distinguished according to the grammatical function of the FIGURE and EFFECTOR arguments and the structural realisation of the predicate expression. Predicate expressions may be realised as simple lexical units (i.e. lexical predicate heads with all dependencies open, cf. *suchen* 'to search') or as (partially) complemented syntagms (i.e. lexical predicate heads that have been compositionally combined with one or more of their complements, e.g. *jemanden durchsuchen* 'to frisk someone'). All lexical complements that are not at the same time arguments of the schematic pattern (i.e. its FIGURE, GROUND or EFFECTOR) are analysed as elements of an "augmented predicate" (i.e. a predicate head in compositional combination with one or more dependents). In contrast to simple predicates ("PRD"), such "augmented predicates" are labelled "PRD+" in the argument structure representation.

For means of illustration, figures (1)–(3) show the representation of three semantically similar, but formally distinct realisations of argument structure patterns with the prepositions *nach*, *um* and *auf*: figure 1 shows an active pattern with two arguments, the marker *nach* and a dative GROUND argument. Figure 2 also shows an active pattern with two arguments, but here with the marker *um* and an accusative GROUND. Figure 3 shows a passive pattern with an augmented predicate and only one realised pattern argument, the accusative GROUND introduced by *auf:*

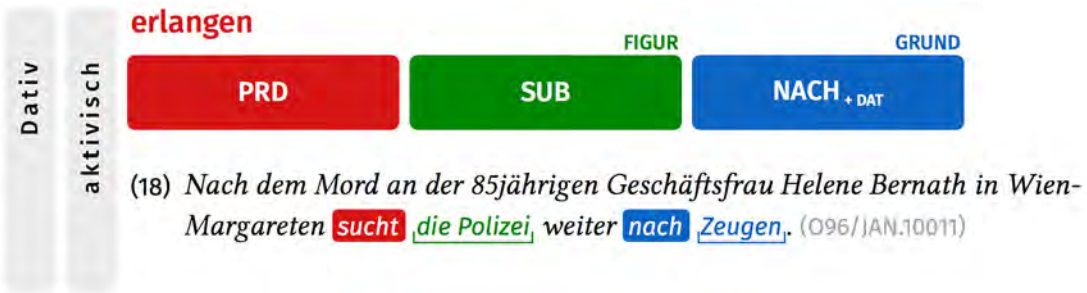


Fig. 1: Sample representation, *nach*-pattern ERLANGEN, with GROUND in dative case

[6] This excludes types that differ from standard realisations only because of construction-independent grammatical regularities (e.g. subjectless variants that owe the missing argument to an infinitival or imperative realisation of the predicate). Instances involving such special constructions ("Sonderformen") can be identified in the pattern search, however.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Fig. 2: Sample representation, *um*-pattern ERHALTEN, with GROUND in accusative case

Fig. 3: Sample representation, *auf*-pattern ERMITTELN, with passive voice and augmented predicate ("PRD+")

3.3 Predicates section

Predicates instantiating a given pattern are listed according to their predicate type (simple non-reflexive verbal predicates, simple reflexive verbal predicates, complex non-reflexive predicates and complex reflexive predicates). The four lists of predicates are accompanied by an annotated corpus example for each element on these lists. In addition to these mere lists, instantiating predicates are grouped into semantic classes, and their interaction with the semantics of the pattern is discussed.

The predicate section closes with a commentary on "shared arguments" that addresses the question of whether the GROUND argument of the pattern is also an argument of particular instantiating predicates. Where this is the case and the lexical argument is also specified as a PP of the relevant type (e.g. *nach* + dative, as in *sich nach etwas erkundigen* 'to ask about something' in the ERLANGEN-pattern), the GROUND argument is considered as shared by the pattern and the instantiating predicate; it is an argument of both.

In other cases, the GROUND argument may also be a semantic argument of the instantiating predicate but is conventionally realised in a different form, for example as an accusative NP (2.a), as a finite complement clause (2.b) or as a PP with a different preposition (2.c):

(2) a. *Er sucht nach einer Wohnung. / Er sucht eine Wohnung.*
'He searches for an apartment.'
b. *Es wird nach dem Sinn von Agentensystemen gefragt. / Es wird gefragt, ob Agentensysteme sinnvoll sind.*
'People wonder about the purpose of agent systems/whether agent systems make sense.'
c. *Die Kinder betteln nach / um Schokolade.*
'The children beg for chocolate.'

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

In case the alternative realisation is in fact the canonical one (as in *betteln um* as opposed to *betteln nach*), attestations in the target construction indicate that this construction is productive, since it encroaches on the semantic territory of a competitor. Where the GROUND argument of the pattern is not an argument of the instantiating predicate, valency approaches analyse it as an adjunct (e. g. *nach etwas buddeln* 'strive to obtain something by digging', *nach etwas kratzen* 'strive to obtain something by scratching' and *nach etwas brüllen* 'strive to obtain something by roaring'). From a pattern-based perspective, however, these arguments are said to be 'coerced' on the predicate by the schematic argument structure pattern, thus testifying to its independent meaningfulness (Goldberg 1995).

4. Outlook: potential applications

Although the resource is primarily intended as a service to the scientific community in linguistics, previous presentations of the project have indicated that it may be worthwhile to showcase specific possible applications of the patternbank also for other user types once the prototype is publicly available.

One such context is academic teaching. Instructors have pointed out that it would be very interesting for them to provide students with resources that permit a direct, side by side comparison of lexical valency- and pattern-based descriptions of one and the same argument structure construction. For predicates that occur in a particular pattern that are also covered in the German electronic valency dictionary E-VALBU,[7] we therefore provide direct links to the relevant verb (reading) entry in E-VALBU.

A second possible application context is second language learning. Governed prepositions pose notorious problems to learners since whatever semantic content they may have retained is not always immediately apparent (cf. section 2.1.1). A description that helps to uncover traces of their original local meaning and puts a given pattern in semantic perspective (by grouping individual expressions with similar uses) may thus be a welcome addition to the spectrum of existing German language teaching materials. Here, it might be worthwhile to consider a gamification component and devise e. g. some kind of quiz that can serve as an entry point into the patternbank.

Last but not least, the predicate lists in individual pattern articles (alongside the associated corpus samples) also perform a quasi-thesaurus function for expressions that instantiate a given grammatical pattern meaning in the German reference corpus DeReKo. Conceivably, this kind of information could also be of interest for other types of linguistic professionals such as translators.

References

DeReKo: Deutsches Referenzkorpus / Archiv der Korpora geschriebener Gegenwartssprache 2022-I-RC1 (Release 08-03-2022). Mannheim: Leibniz-Institut für Deutsche Sprache. www.ids-mannheim.de/DeReKo (last access: 04-05-2022).

Dudenredaktion (no date): Duden online Wörterbuch. https://www.duden.de/woerterbuch (last access: 04-05-2022).

[7] https://grammis.ids-mannheim.de/verbvalenzhttps://grammis.ids-mannheim.de/verbvalenz (last access: 25-03-2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Duden – Die Grammatik (2016): 9th revised and updated edition. Wöllstein, A./Dudenredaktion. Berlin.

Engelen, B. (1975): Untersuchungen zu Satzbauplan und Wortfeld in der geschriebenen Sprache der Gegenwart. München.

E-VALBU: Leibniz-Institut für Deutsche Sprache: „Wörterbuch zur Verbvalenz". Grammatisches Informationssystem grammis. DOI: 10.14618/evalbu: Permalink: https://grammis.ids-mannheim.de/verbvalenz (last access: 04-05-2022).

Goldberg, A. E. (1995): Constructions: a construction grammar approach to argument structure. Chicago/London.

Goldberg, A. E. (2006): Constructions at work: the nature of generalization in language. Oxford.

Helbig, G./Buscha, J. (2017): Deutsche Grammatik: Ein Handbuch für den Ausländerunterricht. [1st edition 2001]. Stuttgart.

Herbst, T. (ed.) (2019): From lexicography to constructicography/Von der Lexikographie zur Konstruktikographie/De la lexicographie à la constructographie. (= International Annual for Lexicography / Revue Internationale de Lexicographie / Internationales Jahrbuch für Lexikographie 35). Berlin/Boston.

Höllein, D. (2019): Präpositionalobjekt vs. Adverbial: Die semantischen Rollen der Präpositionalobjekte. Berlin/Boston.

Kiss, T./Müller, A./Roch, C./ Stadtfeldt, T./Börner, K./Duzy, M. (2016): Ein Handbuch für die Bestimmung und Annotation von Präpositionsbedeutungen im Deutschen. (= Bochumer Linguistische Arbeitsberichte 14). 2nd edition. Universität Bochum. Sprachwissenschaftliches Institut.

Levin, B. (1993): English verb classes and alternations: A preliminary investigation. Chicago/London.

Lyngfelt, B./Borin, L./Ohara, K./Timponi Torrent, T. (eds.) (2018): Constructicography: Constructicon development across languages. Amsterdam/Philadelphia.

Müller, W. (2013): Wörterbuch deutscher Präpositionen. Vol. 1–3. Berlin/Boston.

Perek, F. (2015): Argument structure in usage-based construction grammar: experimental and corpus-based perspectives. (= Constructional Approaches to Language 17). Amsterdam/Philadelphia.

Pittner, K. (1999): Adverbiale im Deutschen: Untersuchungen zu ihrer Stellung und Interpretation. (= Studien zur deutschen Grammatik 60). Tübingen.

Schmitz, W. (1964): Der Gebrauch der deutschen Präpositionen. 3rd revised edition. München.

Schröder, J. (1986): Lexikon deutscher Präpositionen. Leipzig.

Schulz, D./Griesbach, H. (1982): Grammatik der deutschen Sprache. 11th edition. München.

VALBU (2004) = Schumacher, H./Kubczak, J./Schmidt, R./de Ruiter, V. (2004): Valenzwörterbuch deutscher Verben. Tübingen.

VDE (2004) = Herbst, T./Heath, D./Roe, I. F./Götz, D./Klotz, M. (2004): A valency dictionary of English. A corpus-based analysis of the complementation patterns of English verbs, nouns and adjectives.

Weinrich, H. (unter Mitarbeit von Thurmair, M./Breindl, E./Willkop, E.-M.) (2005): Textgrammatik der deutschen Sprache. 3rd revised edition. Hildesheim/Zürich/New York.

Zifonun, G./Hoffmann, L./Strecker, B. (1997): Grammatik der deutschen Sprache. (= Schriften des Instituts für Deutsche Sprache 7.1–7.3). Berlin.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Contact information

Kristel Proost
Leibniz-Institut für Deutsche Sprache (IDS)
proost@ids-mannheim.de

Arne Zeschel
Leibniz-Institut für Deutsche Sprache (IDS)
zeschel@ids-mannheim.de

Frank Michaelis
Leibniz-Institut für Deutsche Sprache (IDS)
michaelis@ids-mannheim.de

Jan Oliver Rüdiger
Leibniz-Institut für Deutsche Sprache (IDS)
ruediger@ids-mannheim.de

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Anna Vacalopoulou/Eleni Efthimiou/
Stavroula-Evita Fotinea/Theodoros Goulas/
Athanasia-Lida Dimou/Kiki Vasilaki

ORGANIZING A BILINGUAL LEXICOGRAPHIC DATABASE WITH THE USE OF WORDNET

Abstract This paper reports on the restructuring of a bilingual (Greek Sign Language, GSL – Modern Greek) lexicographic database with the use of the WordNet semantic and lexical database. The relevant research was carried out by the Institute for Language and Speech Processing (ILSP) / Athena R.C. team within the framework of the European project Easier. The project will produce a framework for intelligent machine translation to bring down language barriers among several spoken/written and sign languages. This paper describes the experience of the ILSP team to contribute to a multilingual repository of signs and their corresponding translations and to organize and enhance a bilingual dictionary (GSL – Modern Greek) as a result of this mapping; this will be the main focus of this paper. The methodology followed relies on the use of WordNet and, more specifically, the Open Multilingual WordNet (OMW) tool to map content in GSL to WordNet synsets.

Keywords WordNet; semantic network; lexicographic database; multimodal database; sign language resources; Greek Sign Language

1. Paper outline

The first section of the paper gives a brief account of all the different parameters of this reorganization process. After a small summary of the Easier project, in the framework of which this research was undertaken (section 1.1), follows a description of the existing lexicographic database, namely, the Noema+ bilingual dictionary, which was based on a multimodal bilingual corpus (section 1.2). The main part of the paper gives an account of the methodology that was used (section 2), first to automatically link the video signs of the dictionary to different synsets in WordNet (section 2.1), as well as to manually review, amend, and validate this mapping (section 2.2.1) through the use of the Greek part of the Open Multilingual WordNet tool. After that, follows an account of the enhancement of the dictionary database based on the semantic and lexical network of relations between words, namely, synonymy, and then between synsets themselves, namely, hyponymy/hypernymy (section 2.2.2). The paper concludes with an evaluation of this semi-automatic enhancement process and some suggestions for future research (section 3).

1.1 The Easier project

Easier (intElligent Automatic SIgn languagE tRanslation)[1] is an ambitious project undertaken by fourteen European institutions that have joined forces to make available, within a period of three years, a unique and innovative service, i. e., an intelligent machine translation framework to bring down language barriers among several spoken/written and sign languages. The main aim of Easier is to develop a service that will facilitate automatic translation between any two pair of European languages, be them sign languages (SL) or spoken/

[1] https://www.project-easier.eu

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

written languages. To this end, project partners contribute their own tools, technologies, and resources (for different sign languages). The relevant technologies that Easier comprises in this process are SL animation, SL recognition, machine translation, and communication technologies.

As far as the translation component is concerned, one of the main challenges has been to gather enough amounts of data for the required training of the machine translation system. Even though most of the European SLs share a good part of their grammar (including their phonology)[2] and some lexicon (cf. e.g., Pizzuto/Volterra 2000), which is utilised in the project, there is great diversity among various sign languages. To make things harder, the raw data which will feed the machine learning process, however rich, is quite diverse. It, therefore, brings together a collection of different video formats, types of material (e.g., corpora, glossaries, dictionaries), annotation schemes and annotation levels (e.g., in the inclusion of SL-specific aspects such as handedness),[3] transcription focus (i.e., phonology as opposed to meaning and function), etc.

In addition, data is scarce for most SLs, even in the EU context, for a variety of reasons. Apart from the fact that all SLs are minority languages and, as a result, they are more likely to be low on resources compared to more widely used languages, collecting and properly annotating SL material is a time-consuming process, which demands a considerable investment in terms of both financial and human capitals (Vacalopoulou 2020, p. 431). In an attempt to bridge this gap, project partners have been linking different SL resources together and providing them with a common detailed phonological representation system. A central part of this effort is the attempt to harmonize this diverse set of data and make them usable for machine learning purposes. To this end, the project team, led in this task by the Institute for German Sign Language and Communication of the Deaf at the University of Hamburg, have been linking different SL resources together and providing them with a common detailed phonological representation system. The goal is to come up with a transferable system of phonological representation and grammar to be utilized for more under-resourced European SLs.

One of the ways to proceed with this connection is the utilization of the WordNet semantic and lexical database. This selection was largely based on the assumption that the translation of European SLs into their respective spoken/written languages could never accurately grasp the full meaning of the original sign content; inevitably, something would be lost in translation. If one was to project this deviation in meaning across all languages of the project (both signed and spoken), it is not hard to understand that a considerable amount of meaning could be lost in the process. Therefore, the WordNet solution was seen as a way to get round the problem of spoken language interference and find a way to connect signs from different SLs via their meaning (see Bigeard et al. 2022).

[2] According to Brentari/Fenlon/Cormier (2018, Summary), "sign language phonology is the abstract grammatical component where primitive structural units are combined to create an infinite number of meaningful utterances. Although the notion of phonology is traditionally based on sound systems, phonology also includes the equivalent component of the grammar in sign languages, because it is tied to the grammatical *organization*, and not to particular content."

[3] Handedness in sign production refers to the dominant hand a signer uses without altering the intended meaning of the content.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

This paper will focus on the experience of the ILSP team with the use of WordNet and, more specifically, the Open Multilingual WordNet (OMW) tool to map content in Greek Sign Language (GSL) to WordNet synsets with a twofold purpose in mind:

1) Contribute to a multilingual repository of signs and their corresponding translations, and
2) Organizing and enhancing the Noema+ bilingual dictionary (GSL – Modern Greek, hereby 'Greek') as a result of this mapping; this will be the main focus of this paper.

1.2 The Noema+ multimodal database

One of the sign resources that have been gathered to contribute to the Easier project is Noema+,[4] a bilingual (Modern Greek – GSL) dictionary currently comprising more than 12,000 entries. This is the most extensive reference work for this language pair to date that has combined various smaller resources and undergone multiple phases of revision and update. The lexical database of the dictionary is a fully annotated multipurpose-multiuse resource.

In terms of its source material, Noema+ was based on the extensive Polytropon bilingual corpus.[5] This resource has been used to build several end products, mainly targeting the bilingual education of deaf children and GSL learning as a second language, as well as services such as the enhancement of the official platform for secondary education in Greece, and an e-class platform as adapted by the Technical Vocational Institute of Athens for accessibility (Efthimiou et al. 2016). Furthermore, it is exploitable for developing a series of SL technologies, including information extraction, Web accessibility tools, incorporation of SL lexical information in natural language processing systems as in the case of machine translation from and into GSL, creation of training material for sign recognition and input to sign synthesis tools enabling signing by virtual signers, i. e., avatars. This extensive corpus was incorporated in the dictionary after an evaluation process of several internal and external stages involving lexicographers, GSL experts, and end-users (Efthimiou et al. 2019). The dictionary was developed in SiS-Builder,[6] a specially designed web-based open environment that enables lexicographers to access other relevant lexical resources and tools in the compilation process (Goulas et al. 2010).

In terms of dictionary microstructure, each video entry consists of one or several translation equivalents in Greek, the use of which is shown in simple, one-sentence examples in both languages. Other microstructure elements contain GSL synonyms, such as *ξυράφι* ('razor') in Figure 1; Greek synonyms, which were added mainly for search purposes to offer a variety of starting points to users who want to look up a certain sign; and multi-word expressions which are cross-referenced to the respective single-word entries.

[4] Noema+ is freely available in ILSP's Sign Language Technologies team website: http://sign.ilsp.gr/signilsp-site/index.php/en/home-3/.

[5] Part of the annotated Polytropon bilingual parallel corpus is freely available via the *clarin:el* repository, the Greek sector of CLARIN, the European infrastructure for language resources and technology: https://inventory.clarin.gr/corpus/835.

[6] SiS-Builder can be accessed in http://sign.ilsp.gr/sisbuilder/.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Fig. 1: GSL synonyms (two different signs) for the same meaning (*ξυράφι*, 'razor') in Noema+

2 Linking Noema+ to the WordNet database

The decision to map the entries in Noema+ to the WordNet database was based on the idea to assign concepts directly to signs without having to go through the words of oral language. Compared to more extended sign resources such as corpora, Noema+ is more context-free as it works primarily as a standalone bilingual dictionary; as in any dictionary, its entries consist of items in isolation – in this case, signs – which are then put into context in the examples of use. Starting from work undertaken for the Easier project, this resource's entries were mapped to corresponding WordNet concepts. The goal was, firstly, to enhance the content of the dictionary database with more alternative translations of signs in Greek and, secondly, to attach translations in more languages towards the future end of making the database multilingual.

2.1 Automatic mapping to the Greek WordNet

In the first stage of this process, the dictionary entries were automatically mapped to WordNet synsets through the use of OMW. Synsets are sets of unordered synonyms that correspond to the same concept accompanied by a simple definition that can have the form of an "explanatory gloss" (Fellbaum 1998) and, in some cases, by a domain label (Fellbaum 2006). An example of the two synsets for the Greek word *βιβλίο* ('book') can be seen in Figure 2. Thus, synset 03165211-n corresponds to 'an accounting journal as a physical object' whereas synset 02870092-n covers 'physical objects consisting of a number of pages bound together'. OMW provides access to synsets in a multitude of languages, linking back to the respective WordNets (Bond/Paik 2012). Figure 3 presents an example of the 02870092-n synset in various languages.

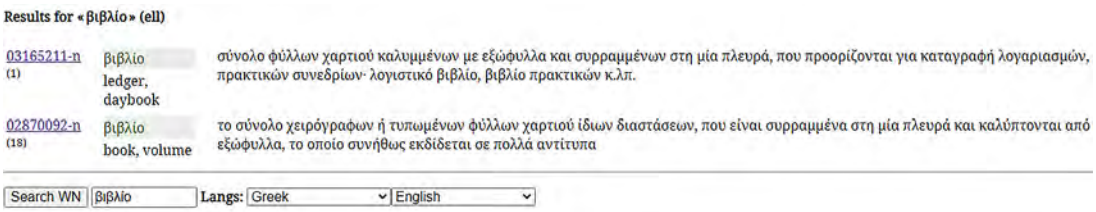


Fig. 2: Search results for the Greek word *βιβλίο* in the Open Multilingual WordNet

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

02870092-n 'physical objects consisting of a number of pages bound together'; Search WN

Albanian	*vëllim , libër*
Arabic	*كِتاب , مُجلّد*
Bulgarian	*издание , книга , том*
Catalan	*llibre , volum , exemplar*
Chinese (simplified)	*书籍 , 书*
Danish	*bog*
Greek	*βιβλίο*
English	*book*$_{10}$ (▸) , *volume*$_{8}$
Basque	*liburu*
Finnish	*volyymi , kirja*
French	*volume , livre*
Galician	*volume , libro , exemplar*
Croatian	*knjiga*
Indonesian	*kitab , jilid , buku*
Icelandic	*bindi , bókarbindi*
Italian	*tomo , volume , libro*
Japanese	*本 , 書 , 冊 , 冊子本 , 冊子 , 図書 , ブック , 篇帙 , 一巻 , -冊 , 書誌 , 書物 , 書籍 , 書史 , 書冊 , 書帙 , 書巻*
Lithuanian	*tomas , knyga*
Dutch	*jaargang*

Fig. 3: Book in different languages in the Open Multilingual WordNet

In this initial phase, the Greek part of WordNet was downloaded from OMW in the form of a text file with tab delimited values, which is available in the rich resources archive of the website. This file comprises approximately 42,200 lines and contains entries that are broken down in several rows. This format was not very practical to use in parallel with the Noema+ database; as a result, the table was converted to a more suitable layout, in which all the contents of the same lexical entry were moved in the same row. Thus, as the goal was to check concepts against GSL video signs without the interference of their Greek equivalents, lexical items sharing a common WordNet ID were grouped together in one entry containing the entire WordNet synset. Figure 4 shows the initial arrangement of the WordNet entry #00006238-v for the Greek *αποχρέμπτομαι* ('expectorate') and the results of the automatic rearrangement in the mapping process. In the first version, the contents of the synset (*αποχρέμπτομαι*, *πτύω*, and *φτύνω*) are aligned horizontally, whereas in the second version, they are arranged vertically to facilitate the automatic mapping process.

Result Grid | Filter Rows: 00006238-v | Edit: | Export/Import: | Wrap Cell Content:

ID	wordnetID	ell:lemma	subID	def
15	00006238-v	αποχρέμπτομαι, πτύω, φτύνω	0	αποβάλλω φλέγματα από τα πνευμόνια με απόπτυση
NULL	NULL	NULL	NULL	NULL

Result Grid | Filter Rows: 00006238-v | Export: | Wrap Cell Content: | Fetch rows:

wordNetID	lemma	subID	Definition
00006238-v	αποχρέμπτομαι		
00006238-v	πτύω		
00006238-v	φτύνω		
00006238-v		0	αποβάλλω φλέγματα από τα πνευμόνια με απόπτυση

Fig. 4: Initial layout (above) and its rearrangement (below) of the OMW entry for *αποχρέμπτομαι* ('expectorate') before mapping it to the Noema+ lexical database

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

In the next stage, the newly created table (WordNetGR) was joined with the bilingual dictionary database with SQL queries. As the entire synset was now one single entry in the table, it was split by the number of its contents within the SQL query so that it could match dictionary entries and thus create the mappings to the GSL videos. An example can be seen in Figure 5, which shows WordNet ID #00220869-v for *δυναμώνω* ('strengthen'), which has 2 delimiters as the synset contains 3 Greek synonyms: *δυναμώνω*, *ενισχύω*, *ισχυροποιώ*. In order to execute the SQL join query, it was necessary to split the field to its contents. The corresponding dictionary entry to which this synset was mapped was rearranged to a total of three entries, each containing one of the three Greek synonyms. The third entry was included in the SQL-results, because the corresponding equivalent in Greek is not contained in the bilingual dictionary.

The result of this process was the full mapping of dictionary entries to corresponding synsets in WordNet. For quality control purposes, this result was later double-checked by GSL experts, who made corrections if appropriate.

Result Grid | Filter Rows: | Export: | Wrap Cell Content: | Fetch rows:

wordNetID	ell:lemma	Ελληνικό αντίστοιχο (NOHMA)	ell:lemma1	ell:lemma2	ell:lemma3	ell:lemma4
00220869-v	δυναμώνω, ενισχύω, ισχυροποιώ	δυναμώνω	δυναμώνω	ενισχύω	ισχυροποιώ	NULL
00220869-v	δυναμώνω, ενισχύω, ισχυροποιώ	ενισχύω	δυναμώνω	ενισχύω	ισχυροποιώ	NULL
00217499-n	δυστύχημα, καταστροφή, συμφορά	δυστύχημα	δυστύχημα	καταστροφή	συμφορά	NULL
00217499-n	δυστύχημα, καταστροφή, συμφορά	καταστροφή	δυστύχημα	καταστροφή	συμφορά	NULL
00212414-v	διατηρώ, προφυλάσσω, συντηρώ	διατηρώ	διατηρώ	προφυλάσσω	συντηρώ	NULL
00212414-v	διατηρώ, προφυλάσσω, συντηρώ	προφυλάσσω	διατηρώ	προφυλάσσω	συντηρώ	NULL

Fig. 5: WordNet ID #00220869-v for *δυναμώνω* ('*strengthen*') produced two extra entries in the dictionary database after mapping the entries to the corresponding synonyms

2.2 Manual organization of the Noema+ database

The second part of the mapping process was the manual processing of the automatically generated results. The purpose of this stage was to review and amend these results on one hand, and to enhance the content of the dictionary database on the other hand.

2.2.1 Review and amendments

Following the automatic mapping of GSL signs to their corresponding Greek words in OMW, an important goal was to evaluate whether GSL signs and OMW Greek words were semantically equivalent. To this end, each sign was carefully checked by expert GSL signers against its corresponding OMW synset to ensure that they shared the same meaning. In order to decide whether an OMW Greek equivalent indeed corresponded to the meaning of a particular sign, several parameters were taken into consideration including the videos of both isolated signs and their linked examples of use in the Noema+ dictionary, as well as the definitions or explanatory glosses of synsets. An example of this process would be the sign entry for *αστέρι* ('star'), which had been automatically linked to the following two Greek OMW synsets:

1) WordNet ID #09444100-n, i. e., "a celestial body of hot gases that radiates energy derived from thermonuclear reactions in the interior", and
2) WordNet ID #09762509-n "someone who is dazzlingly skilled in any field".

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Although the mapping was successful in terms of the equivalents in Greek, it is only the first sense of this polysemous word that corresponds to the video equivalent found in the GSL database. The second sense was, therefore, deleted from the dictionary database. Luckily, almost 70% of the GSL synonyms that had been automatically linked to the same WordNet ID were verified as correct matches at this stage, rendering the automatic mapping process successful.

Following the abovementioned process, all entries in the new combined database were double-checked manually one by one by GSL experts, and items were deleted or verified accordingly, before proceeding to the next stage of dictionary enhancement.

2.2.2 Enhancing the Noema+ database through WordNet

The next, and final, stage in this process was the enhancement of the dictionary database with new entries as a result of their mapping to WordNet synsets. This was pursued mainly by exploiting the richness of each synset, which in some cases contained multiple synonymic items, and by further experimenting with other lexical relations, such as hyponymy and hypernymy.

In regard to the synsets, it was found that they were helpful in the enrichment of the dictionary with multiple Greek equivalents in this relatively quick, semi-automatic process. Thus, more synonymic items from the corresponding synset were added to the database under each GSL entry after being validated by GSL experts. Examples of this type of enrichment include items such as *κατάστημα* ('store') and *μαγαζί* ('shop'), or *αναζητώ* ('search') and *ψάχνω* ('look for').

Apart from this first-level enhancement, the mapped WordNet synsets gave the team of experts the opportunity to further explore polysemy in the context of GSL by adding more senses to specific video signs. For instance, whereas the sign for *άνεση* ('comfort') in GSL had been automatically mapped to WordNet ID #14491889-n, i.e., "freedom from financial difficulty that promotes a comfortable state", careful examination of the Greek WordNet in OMW produced additional senses of the corresponding word in Greek that were found to be translations of the original GSL sign as well. In this case, as shown in Figure 6, the same video was linked to two additional WordNet IDs, namely, #14445379-n ("a state of being relaxed and feeling no pain"), and #07492516-n ("a feeling of freedom from worry or disappointment"), as all three senses are represented by the same sign in GSL. The added value of this procedure was that it allowed for a more detailed documentation of polysemy in GSL.

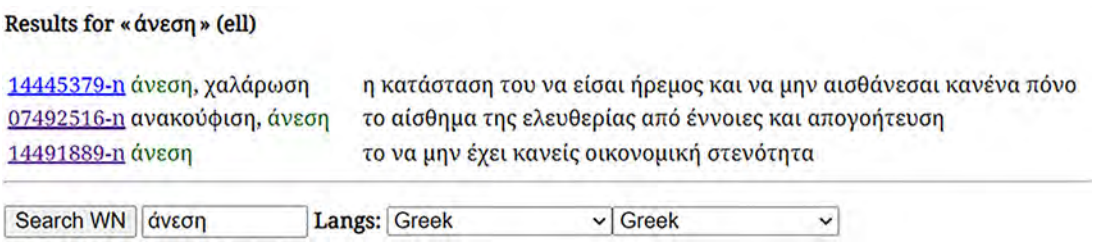


Fig. 6: Multiple WordNet IDs were linked to the same GSL video sign allowing for a more detailed documentation of GSL polysemy

Taking the enhancement process one step further from the synset itself, effort was made to explore lexical relations in which synsets themselves are connected to one another, namely,

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

hyponymy and hypernymy. Thus, synsets that link to other synsets through such relations were analysed for the possibility of offering ways to further enhance the dictionary database.

Fig. 7: Lexical relations of synset #10794014-n for *συγγραφέας* ('writer') with other synsets

A relevant example is the case of *συγγραφέας* ('writer'), which is connected to multiple hyponyms as shown in Figure 7. As expected, 'writer' is linked to several hyponymic synsets such as 'biographer', 'dramatist', and 'poet'. As observed, this could be a valuable source of creating additional dictionary entries for senses that had not been linked to a dedicated sign by the natural GSL signers, who served as informers in the development of the corpus upon which the first version of the dictionary was based. An example of this would be the hyponym of 'writer' δραματουργός ('playwright') for which there was no entry in the dictionary database as no dedicated sign had been reported to date. In this case, the lexical relations between synsets recorded in WordNet allowed GSL experts to add a compound sign consisting of the simple signs for *θεατρικό* 'play' and *γράφω* 'write' to represent this concept (Figure 8). Indeed, this kind of compounding is a very common technique in sign language morphology (Sandler/Lillo-Martin 2006, pp. 72–75) and a natural way of producing new lexicon within signing communities. Nevertheless, as valuable, and extremely promising as this technique is, it needs to be applied with extra care, as its results are far from self-evident and require multiple levels of validation by native signers and/or against GSL corpora.

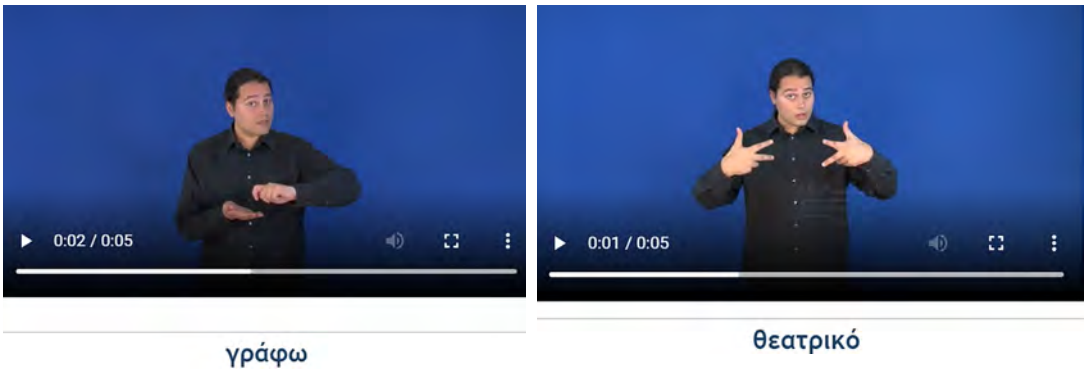


Fig. 8: Combining the signs for *γράφω* (write) and *θεατρικό* (play) to enhance the dictionary database with a new sign for *δραματουργός* (playwright)

3 Conclusion and future steps

This contribution presented an account of the rearrangement and enrichment of the database of the NOEMA+ bilingual GSL – Greek dictionary after mapping its entries to corresponding WordNet synsets. In this ongoing process, more than 1,200 lexical items have already been identified as possible candidates for dictionary enhancement either from analysing the synsets themselves, or by an extended investigation of lexical relations between synsets. These will be evaluated by the team of lexicographers and GSL experts for inclu-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

sion in NOEMA+ as this research is still in progress. Based on these promising results, the team will explore further enhancement possibilities towards opening the dictionary to more languages through their respective WordNets with a view of making the resource multilingual.

References

Bigeard, S./Schulder, M./Kopf, M./Hanke, T./Vasilaki, K./Vacalopoulou, A./Goulas, T./Dimou, A.-L./Fotinea, S.-E./Efthimiou, E. (2022): Introducing sign languages to a multilingual WordNet: bootstrapping corpora and lexical resources of Greek Sign Language and German Sign Language. In: Proceedings of 10th Workshop on the Representation and Processing of Sign Languages: Multilingual Sign Language Resources (LREC-2022), Marseille, France, 20–25 June 2022.

Bond, F./Paik, K. (2012): A survey of wordnets and their licenses. In: Proceedings of the 6th Global WordNet Conference (GWC 2012). Matsue, pp. 64–71.

Brentari, D./Fenlon, J./Cormier, K. (2018): Sign language phonology: Oxford research encyclopedia of linguistics. https://oxfordre.com/linguistics/view/10.1093/acrefore/9780199384655.001.0001/acrefore-9780199384655-e-117#acrefore-9780199384655-e-117-div1-1 (last access: 14-05-2022).

Efthimiou, E./Fotinea, S-E./Goulas, T./Vacalopoulou, A./Vasilaki, K./Dimou, A.-L. (2019): Sign language technologies and the critical role of SL resources in view of future internet accessibility services. In: Technologies 7 (1), 18, pp. 1–21.

Efthimiou, E./Fotinea, S.-E./Dimou, A.-L./Goulas, T./Karioris, P./Vasilaki, K./Vacalopoulou, A./Pissaris, M./Korakakis. D. (2016): From a sign lexical database to an SL golden corpus – the POLYTROPON SL resource. In Proceedings of 7th Workshop on the Representation and Processing of Sign Languages: Corpus Mining (LREC-2016), Portorož, Slovenia, 23–28 May 2016, pp. 63–68.

Fellbaum, C. (ed.) (1998): Wordnet: an electronic lexical database. Cambridge, MA.

Fellbaum, C. (2006): WordNet(s). In: Brown, Keith et al. (eds.): Encyclopedia of language and linguistics. Second Edition. Oxford, pp. 665–670.

Goulas, T./Fotinea, S.-E./Efthimiou, E./Pissaris, M. (2010): SiS-Builder: a sign synthesis support tool. In: Proceedings of the 4th Workshop on the Representation and Processing of Sign Languages: Corpora and Sign Language Technologies (CSLT 2010), Satellite Workshop of the LREC-2010 Conference, Valetta, Malta, 17–23 May 2010, pp. 102–105.

Grigoriadou, M./Kornilakis, H./Galiotou, E./Stamou, S./Papakitsos, E. (2004): The software infrastructure for the development and validation of the Greek Wordnet. In: Romanian Journal of Information Science and Technology 7 (1–2), pp. 89–105.

Noema+ (2022): http://sign.ilsp.gr/signilsp-site/index.php/en/home-3/ (last access: 09-03-2022).

Open Multilingual Wordnet (OMW): http://compling.hss.ntu.edu.sg/omw/ (last access: 14-03-2022).

Pizzuto, E./Volterra, V. (2000): Iconicity and transparency in sign languages: a cross-linguistic cross-cultural view. In: Emmorey, K./Lane, H. (eds.): The signs of language revisited: an anthology to honor Ursula Bellugi and Edward Klima. Mahwah, pp. 229–250.

Polytropon Parallel Corpus, Version: 1.0.0 (2022): https://inventory.clarin.gr/corpus/835 (last access: 21-03-2022).

Princeton University (2010): About WordNet. https://wordnet.princeton.edu/ (last access: 07-03-2022).

Sandler, W./Lillo-Martin, D. (2006): Sign language and linguistic universals. Cambridge.

SiS-Builder (Sign Synthesis Builder) (2022): http://sign.ilsp.gr/sisbuilder/ (last access: 21-03-2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Vacalopoulou, A. (2020): Sign language corpora and dictionaries: a multidimensional challenge. In: Proceedings of the EURALEX XIX Congress of the European Association for Lexicography – Lexicography for inclusion, Alexandroupolis, 7–11 September 2021. Proceedings Book Volume 1, pp. 427–434.

Contact information

Anna Vacalopoulou
Institute for Language and Speech Processing, ATHENA RC
avacalop@athenarc.gr

Eleni Efthimiou
Institute for Language and Speech Processing, ATHENA RC
eleni_e@athenarc.gr

Stavroula-Evita Fotinea
Institute for Language and Speech Processing, ATHENA RC
evita@athenarc.gr

Theodoros Goulas
Institute for Language and Speech Processing, ATHENA RC
tgoulas@athenarc.gr

Athanasia-Lida Dimou
Institute for Language and Speech Processing, ATHENA RC
ndimou@athenarc.gr

Kiki Vasilaki
Institute for Language and Speech Processing, ATHENA RC
kvasilaki@athenarc.gr

Acknowledgements

This work is supported in part by the EASIER (Intelligent Automatic Sign Language Translation) Project. EASIER has received funding from the European Union's Horizon 2020 research and innovation programme, grant agreement n° 101016982.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Dictionary Writing Systems and Lexicographic Tools

XX EURALEX INTERNATIONAL CONGRESS
DICTIONARIES AND SOCIETY
12-16 July 2022, Mannheim, Germany

Nico Dorn

AN AUTOMATED CLUSTER CONSTRUCTOR FOR A NARRATED DICTIONARY

The Cross-reference Clusters of *Wortgeschichte digital*

Abstract *Wortgeschichte digital* (Digital Word History) is an emerging historical dictionary of the German language that focuses on describing semantic shifts from about 1600 through today. This article provides deeper insight into the dictionary's "cross-reference clusters," one of its software tools that performs visualization of its reference network. Hence, the clusters are a part of the project's macrostructure. They serve as both a means for users to find entries of interest and a tool to elucidate relations among dictionary entries. Rather than delve into technical aspects, this article focuses on the applied logics of the software and discusses the approach in light of the dictionary's microstructure. The article concludes with some considerations about the clusters' advantages and limitations.

Keywords Historical lexicography; dictionary; word history; visualization; digital humanities; graph theory

1. Introduction: some characteristics of *Wortgeschichte digital*

Wortgeschichte digital (Digital Word History, or hereafter, WGd)[1] is a digital, monolingual dictionary of the German language that aims to describe the German vocabulary from about 1600 through today, focusing primarily on semantic shifts. A hallmark of the project is the narrative form of its entries, which describe, reflect, and historically contextualize observed developments. This distinguishes it from the majority of other dictionaries available. But WGd does in some respects follow the tradition of historical dictionaries nevertheless, enriching its entries, for example, with an appropriate amount of quotations to illustrate and provide evidence for historical usage. The project began from scratch in 2019; it is still in its early stages, and does not intend to describe the whole of German vocabulary (which would be beyond its personnel resources anyway). Instead, it aims to provide a selection of several thousand lemmas that belong to an array of different topic domains, such as *politics and society, economy,* and *music and arts.* As the dictionary explicitly addresses the public at large, its entries are written in a more relaxed style but without neglecting their scientific nature.

Given the broad target audience, the project strives to enhance user experience by implementing a host of technical tools to its website. Among the already available features are a timeline (*Zeitstrahl*) that serves as an alternative access point to the dictionary's lemmas; tightly integrated help windows that elucidate the usage and meaning of linguistic terms; a quotation navigator (*Belegnavigator*) that provides an overview of the quotations' chrono-

[1] https://wortgeschichten.zdl.org/.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

logical distribution and the spread of historical word forms; and last but not least the cross-reference clusters (*Verweiscluster*) with a clickable, structured lemma list.[2]

Fig. 1: A cross-reference cluster as displayed in the entry Elite; the lemmas are clickable and direct the user to the appropriate entry

This article provides a deeper insight into the cross-reference clusters. In section 2, it delivers an account of the rules devised to obtain a cluster (as shown in Fig. 1). The section not only describes the applied rules but also discusses their underlying logic. Section 3 summarizes the advantages and limitations of this approach with a view toward other projects. First, though, we make some remarks on the microstructure of the WGd dictionary. The considerations in section 2 and the notes on the cross-reference clusters' limitations would otherwise not be comprehensible.

* * *

There are two core ideas or prerequisites on which the cross-reference clusters are based: First, they should allow users to find entries of interest within the dictionary. They should not have to rely on an alphabetical list as the sole entry point and search option. Therefore, from a visual perspective, principles of simplicity and perspicuity must be heeded. Second, the construction of the clusters must be fully automatic. This is of particular importance, as the WGd entries are published continuously rather than in installments, as is usually the case with print publications. That is why the form and content of each cluster are highly dynamic. Since the continuous publication of new entries constantly increases the complexity of the entangled web that links entries with one another, and since the clusters result from an analysis of that very network structure, they have to be redrawn every time a new entry hits the web. This creates a pressing situation, as the continuous publication also leads

[2] For the timeline, see https://www.zdl.org/wb/wortgeschichten/#Zeitstrahl. For the help windows, see, e.g., "Spezialisierung" in the first paragraph at the page https://www.zdl.org/wb/wortgeschichten/Masse. The quotation navigator can be found in every entry. Hit the navigator icon beside the heading "Belegauswahl."

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

to the addition of references in already-published entries, especially when older entries mention a lemma that is the headword of a newly published entry.[3]

Harm (in this volume) provides an in-depth description of the entry's microstructure. But to grasp the conditions that mold the clusters, some key features of the WGd entries need to be outlined as well.

Each WGd entry opens with a summary, followed by an orienting section. In addition to a table of contents, the orienting section incorporates some aspects that usually form the basis of a historical dictionary: listed word meanings and more or less extended word lists of up to three categories: word formations (*Wortbildungen*), word combinations (*Wortverbindungen*), and similar expressions (*bedeutungsverwandte Ausdrücke*). Especially the words listed in the latter category are enriched with details about semantic relations, such as contrast or synonymy. The orienting section precedes the core of every WGd entry: a continuous text that charts the semantic development that the headword underwent within the period under investigation. Many entries have initially hidden text passages that give users in-depth information about a discussed fact. A good example is in the entry *Erika Mustermann* ("Jane Doe"). Its so-called *Mehr erfahren* ("learn more") section outlines the adoption of new identity cards in Germany around 1980. In doing so, it provides some information as to why the authorities resorted to the word *Mustermann* ("average person," with the archaic connotation "wholesome person") for the cards' mock-ups that still circulate today.

In addition to standard entries, the dictionary consists of a small but growing number of overview articles that deal with a whole word field at once.[4] The microstructure of these entries does not vary significantly from the standard layout. They mainly differ in their much broader point of view on semantic developments, which entails some changes to the orienting section. In those entries, enumerated meanings do not make much sense as the text deals only with lemmas for which there are standard entries to provide a much deeper insight.

The situation is further complicated by the fact that the logic of "one entry, one lemma" does not apply to WGd. In many cases, on the contrary, it is of particular importance to describe the semantic change of several lemmas in a sole entry. A striking example is *Beaumonde/die schöne Welt,* where the latter lemma ("the beautiful world") is a loan translation of the first. Hence, the description of both is inextricably entwined. Entries of that type are quite frequent. Furthermore, entries may also have subordinate lemmas (*Nebenlemmata*) that are defined and described much like the main headwords, albeit less comprehensively. The entry *alternativ*, for example, also covers the multi-word unit *alternative Fakten* ("alternative facts"), an expression in which *alternativ* adopts the meaning "bogus, false".

References to different WGd entries can be found in every part of an entry, and they may point to either a main or a subordinate lemma. The cluster constructor evaluates all references regardless of their position in the entry, which can pose some issues.

[3] The lexicographers have a quality assurance tool (cf. section 2.2) that scans all entries once there is a newly published. It specifically tracks terms that are enclosed in special markup code (TEI `<mentioned>`) and external references to entries of our partner project (https://www.dwds.de/). A message is printed if the headword of a new entry matches one of those terms. The tool also ensures markup consistency, such as for diasystemic values and semantic relations. The *Svensk ordbok* relies on a somewhat comparable tool that visualizes cross-references to obliterate errors and gaps (Blensenius et al. 2021).

[4] E. g. https://www.zdl.org/wb/wortgeschichten/Wortfeld-Lebensformen.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Regarding multi-word entries, it is impossible to tell which lemma is exactly the referrer that points to a different lemma. In such a case, we have to resort to the assumption "all headwords point to the referred lemma". The underlying decision to write a multi-word entry is always based on the observation that the headwords need to be described in close conjunction to elucidate their semantic shift and/or historical distribution. We discuss together what belongs together. That is why it would be a difficult thing to partition such an entry into sections that deal solely with one of the lemmas. Besides, such a solution had to impose formal restrictions to the writing process that are not eligible in the light of the project's knowledge interest, even if they were preferable from a technical point of view.

Fortunately, the inverse case does not pose any problems. We can always be sure of which main lemma a reference points to. The same is true for references to subordinate lemmas, as those are always associated with a certain text position from where their description starts.

As mentioned before, some references have a semantic description attached to them. But that is not always the case, again given the specific structure and contents of WGd entries. It would be cumbersome and difficult to attach a specific semantic to every single reference, as there are several reasons why they were added, and these reasons exceed baseline categories like synonymy or hypernymy by far: There are references that point to lemmas with a comparable semantic shift; others refer to lemmas that are part of the same word family; others point to the descriptions of a historical context that is given in a different entry and so on. Not all of these are linguistic categories. It would be quite difficult to tackle this issue, especially with limited personnel resources, but a denser markup would clearly have computational advantages (cf. Meyer/Müller-Spitzer 2010).

Again, all these issues arise mainly because WGd entries are written as a continuous text that includes information that exceed or differ largely from typical lexical resources. However, the approach chosen for the cluster constructor alleviates those issues substantially.

2. On constructing a cross-reference cluster

2.1 Basic rules

The construction of the cross-reference clusters is not as dependent on advanced programming skills as one might think. The usage of some basic logic and a limited understanding of a graph structure with directed edges is much more important to achieve the outcome as shown in Figure 1. When the baseline situation is as intricate as outlined in section 1, a good first step is to prune every distracting factor. Therefore, as a first step, we temporarily leave aside all the difficulties that the complex structure of the WGd entries imposes on us. For the time being, the references will be stripped to the core, regarded as if no semantic relation were attached to them at all. Thus, they have no particular remarkable quality other than pointing. This implies of course that they all have the same weight, which means none has conspicuous importance, a condition that enables us to operate on them with a very basic logic that will lead to a clear, programmable solution.

When the quality of all references is identical, it is a valid approach to reduce the issue in a way that only three relational types remain:

1) A $\rightarrow$ B (A points to B)

2) B $\rightarrow$ A (B points to A)

3) A $\leftrightarrow$ B (A points to B *while* B points to A)

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

The third type looks promising, especially when this relation pertains to more than two entities at once; in that case, the entities refer so densely to one another that they are virtually trapped in a reciprocal structure. That is an interesting signal of proximity. We cannot say anything about the quality of this proximity (as we stripped the references of the distinguishing meanings they might have had), but we are able to posit a first rule:

(1) Every time we encounter a reciprocal reference structure, all the involved lemmas belong to a cluster center (*Clusterzentrum*) when the lemmas are also distributed over at least two different entries.

The lemmas *Elite, elitär* ("elitist"), and *Establishment,* in the dark gray area of the example cluster in Figure 1, abide by this rule because they exhibit a reference structure like this:

Elite ↔ *elitär*
Elite ↔ *Establishment*
elitär ↔ *Establishment*

In other words, every lemma of a cluster center points to every other lemma that belongs to the same center (otherwise they would not form a cluster center). What we see is a state of maximal reciprocity.[5] But if we had left it at that, we would have discarded a host of references because such a tight reference structure, as it can be observed within a cluster center, is the exception rather than the rule. Thus, to include at least some of the missed references we deploy another rule:

(2) Every time a lemma that belongs to a cluster center points to another lemma that does not belong to the same center, the latter lemma is part of the cluster fringe (*Clustersaum*).

In the example cluster of Figure 1, the fringe encompasses all the lemmas in the light gray area, from *High Society* to *Verschwörungstheorie* ("conspiracy theory"). Again, the assumption behind the rule is that every reference is of equal importance. Therefore, when a lemma does not belong to a cluster center but is referred to from within that center, it will be closely related to it. That is even more the case if the particular lemma points back to a lemma within the center while that very lemma points to the fringe lemma. (If the fringe lemma had pointed to *all* the lemmas in the cluster center, it would have formed part of the center itself. But this can be ruled out since we checked that before.) This observation leads us to the next rule:

(3) Every time a lemma that belongs to the cluster fringe forms a reciprocal reference with a lemma within the cluster center, this lemma is more pertinent to the lemmas in the center than to the other lemmas at the fringe.

In the example cluster, this is the case for *High Society* and *Masse* ("mass") but not for *alternativ* and *Verschwörungstheorie.*[6] Thus, the first two are promoted and bunched into a distinctive bundle (*Bündel*), and as a consequence are printed above the latter lemmas in a larger font.

The next rule leads to the inclusion of all the lemmas against the white background, from *Jetset* to *Kaste* ("caste"). In our initial considerations, it follows the second relational type, at least on the assumption that entity A were a lemma of the cluster center and entity B a lem-

[5] This is a plain example because each lemma has its own entry. That is not always the case (see section 1).

[6] In fact, *High Society* points to *Establishment* and vice versa. The same is true for *Masse* and *Elite.* But this cannot be deduced from the cluster itself; one must consult the source code or the quality assurance tool (cf. fn. 3).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

ma that did not pertain to that center (in that sense, the first relational type is already covered by rule 2):

(4) A lemma that points to a lemma of a cluster center belongs to the cluster periphery (*Clusterumfeld*) when no lemma of the cluster center points back to that very lemma from the periphery.

To be precise, the final clause of this rule is superfluous and only for clarity; for if a lemma from the cluster center points to another lemma, rule 2 is applicable, and that rule excludes a lemma automatically from being part of the cluster fringe because:

(5) A lemma must not appear in more than one of the three cluster circles (*Clusterkreise*).

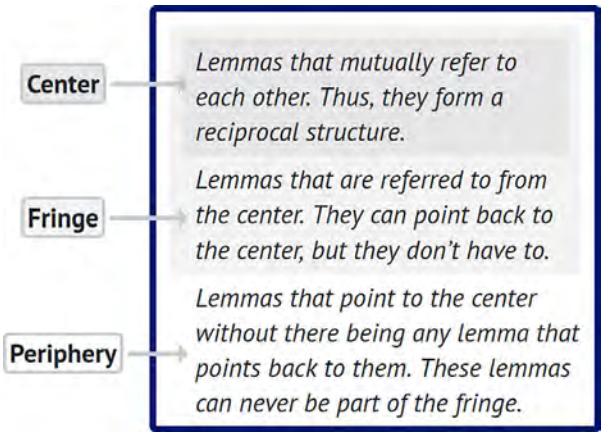


Fig. 2: The basic structure of a cross-reference cluster

What results from these considerations is the basic structure of a cross-reference cluster (Fig. 2). This consists of three circles (center, fringe, periphery), whereas the fringe can be subdivided into two distinct bundles. It is important to reiterate that the resulting clusters are all constructed starting from the cluster center. From there on, the constructor follows its way down to fill in around the center. Thus, the starting point is an area in which the lemmas are closely related to one another. From there on, we proceed to include much more loosely affiliated lemmas. What all cluster lemmas have in common is a straightforward, direct connection to different lemmas in the center.

2.2 Adding further details

The next step is to enrich the clusters with further details we can obtain from the entries. For one thing, we know the position of a reference in the microstructure of a WGd entry. It is quite simple to recognize whether a reference is rather prominent (e. g., in the summary) or marginal (e. g. in a footnote). We also know about their degree of systematicity. A back reference that points to an overview article or the inclusion of a reference in a structured word list certainly signals that an author intends to systematize an observation. Those references should bear more weight than others. That is why we devised a point system, in order to attach weight to the lemmas and promote those that were referred to more often, more prominently, and/or with a greater degree of systematicity.[7] The points a lemma col-

[7] There are six different systematical positions where references can be found: The entry header may include a reference to a superordinate overview article (if there is one). For this is a quite prominent

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

lects enable us to sort them by weight within their appropriate circle. What results are non-perspectivized clusters (i. e., clusters in their standard shape, see below), as they can be seen on the project's overview page (Fig. 3).

Fig. 3: Two non-perspectivized clusters as they can be seen on the overview page

The word *Lebensformen* ("lifestyles") in the second cluster of Figure 3 is printed in small capitals because it leads to an overview article that deals with the five lemmas that complete the cluster center. There is no specific rule that overview articles should be the first words of a cluster center, but the point system ensures that this is always the case. As every entry that pertains to a word field points back to the superordinate overview article, those references yield a lot of points to the field article, which promotes it in a way that it becomes the cluster's head.

position with a high degree of systematicity, such a reference yields 10 points. Thus, an overview article is always the sole head of its cluster. References in the word lists, which can be found in the orienting section, also exhibit high systematicity and therefore yield 3 p. Sometimes references can be found in the entry's summary (3 p.), a prominent position, but they are usually only part of the main text (2 p.). These 2 p. are something of a baseline on which the whole point system rests. Hence, references in initially hidden passages that provide further information are demoted, as are those in footnotes (1 p.). The subordinate lemmas pose a special problem, as there is no way of telling whether a reference comes from them or the article's main lemmas (see section 1). Therefore, it is assumed that an outgoing reference is only to the main lemmas. Fortunately, the inverse does not pose problems. We can reliably determine whether a reference points to a subordinate lemma so that the point system is not confounded by them. Finally, lemmas within the cluster fringe receive a 1000 p. bonus for every reciprocal reference they form with a lemma of the cluster center (see rule 3). This high number serves as a reliable marker, indicating that the lemmas pertain to the upper bundle of the fringe. Such a high value is never reached by references. The reason we initially applied the point system was to ensure that the superordinate overview articles are always the first lemma in a non-perspectivized cluster; thus the high score for references that point to them. Although the point system is a bit arbitrary, it ensures that lemmas that are referred to more often, more systematically, and more prominently are ranked higher than others.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Now that we can gauge the importance of a word in a cluster circle, we can formulate a threshold condition that splits the cluster center into separate bundles similar to the cluster fringe:

(6) When some lemmas of a cluster center gain significantly more points than the rest,[8] they are more pertinent and therefore promoted.

Thus, those lemmas are printed in the first position, with a gap below and in a larger font. This is the case for *Lebensformen* and *Elite* in Figure 3.

The division of the cluster center into separate bundles is dismissed when a cluster gets perspectivized (as in Fig. 1). In comparing the clusters in Figure 1 and Figure 3, you will notice that they are virtually the same. In actuality, they are the same because the two visualizations rely on the same data.[9] The only difference is that the entry lemma has been printed into an ellipse and a two-sided arrow has been added to indicate the reciprocal relation of the lemmas within the cluster center.

But there is another remarkable change. As the clusters in the entries are perspectivized, we are able to add back the semantic relations we initially dropped (see section 2.1). This step is possible only in a perspectivized cluster, as lexical relations like hypernymy and homonymy, meronymy and holonymy require a certain perspective in the form of a reference lemma to be valid declarations. Every cluster lemma printed in italics has a lexical relation attached to it. These are visible as a tooltip when the mouse hovers over them. Alternatively, the lemmas can be filtered by relation when one clicks the filter icon in the upper right corner (Fig. 4).

Fig. 4: A filtered cluster that hides every lemma unless it is a synonym to the headword

[8] The threshold is derived from experience. While arbitrary, it works out well. It amounts to ≥ 10 % of the first lemma's points but at least 3 p. which is the score a prominent reference yields. As it is applied while the clusters are visualized, it is not present in the data itself.

[9] The data file is publicly available: https://www.zdl.org/static/wortgeschichten/Artikel.json.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

2.3 Lemmas with multiple assignments

As said before, the clusters that result from the application of the given rules are based on the entire cross-reference network that pervades the WGd entries. That is why their default shape is a non-perspectivized cluster. Hence, the clusters are derived not from a word perspective but from the word fields our lexicographers are working on. During the writing process, we do not follow the alphabet but operate within thematic boundaries. That is a condition for the cluster constructor's analyses to produce meaningful results even if there is a relatively small number of published entries.

What results from the fact that clusters are based on word field is that the lemmas have no special allegiance to a particular cluster. They can appear in more than one at the same time, even lemmas that are part of a cluster center. In this respect, a certain fuzziness is observed across the clusters. As we have seen, *Elite, Establishment,* and *elitär* form a cluster center, but *Elite* and *Masse* form a thematically adjacent, alternative center, as the two words exhibit a reciprocal reference structure as well. We do not take that to be an issue. Despite the multiple assignment of lemmas to several clusters the 187 lemmas that have been published to date form only thirty clusters.

Nevertheless, the lexicographers have access to a quality assurance tool that shows them a way to reduce the number of clusters, in addition to coding errors and suggestions for further references when applicable. This tool compares the clusters, lists them, and displays their degree of similarity. If a lexicographer deems it worthy to merge two or more clusters, the tool tells him or her which references need to be added to an entry in order to accomplish the goal.

Merging smaller clusters into larger ones is the only viable option. The underlying program executes its rules thoroughly, which makes it very hard to mold a cluster into a shape of one's preference. We do not think that this poses a problem, for it was the lexicographer in the first place who decided to interlink two entries by adding a reference. The computer only visualizes the results of that decision and helps to keep entries in good shape. It also points to similarities that might have been overlooked. Yet the decision of whether a reference should be added, is not conferred to a computational device—and should not be. Nevertheless, the computers' ability to visualize data is indispensable given that we are dealing with an overload of information (cf. Therón et al. 2014). Visualizations like the cross-reference clusters will definitely alleviate this situation, as one of their strengths is rating of pertinent information.

A graph of the example cluster with directed edges not only illustrates the intricacy of the reference network that pervades the WGd articles (Fig. 5) but also clarifies that, even if the number of vertices is low, the entanglement of the cross-references is at best hard to understand, even when the vertices are color-coded. Nevertheless, such a graph can be helpful, as it highlights the earlier-mentioned fact that lemmas may appear in more than one cluster center. The dark-blue-rimmed lemmas form the center of the example cluster in Figure 1 (*Elite, elitär, Establishment*). Additionally, nodes 1 and 5 (*Elite, Masse*), nodes 3 and 4 (*Establishment, High Society*), and nodes 4, 8, and 9 (*High Society, Jetset, Crème*) form different centers, as their lemmas also refer to one another reciprocally. Furthermore, it is clear that the nodes 2, 5, 6, and 7 (*elitär, Masse, alternativ, Verschwörungstheorie*) are not part of the cluster formed by nodes 4, 8, and 9, as there is no edge that directly connects those lemmas.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

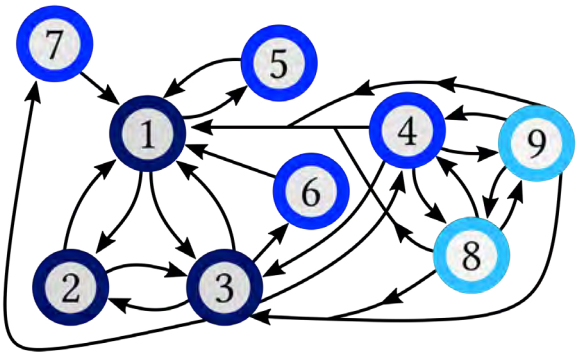


Fig. 5: A directed graph that displays all the references between the first 9 lemmas of Figure 1, whereby vertex 1 is *Elite* and vertex 9 *Crème*

Figure 5 also points to the fact that the construction of cross-reference clusters is essentially a mathematical issue that pertains to graph theory. In that respect, we can reformulate the problem of finding cluster centers as a variant of a clique problem. What the cluster constructor actually does is find all cliques in a graph (like the one shown in Fig. 5), whereas a clique is a subgraph that is complete in the sense that all vertices of the subset are adjacent. That means that all vertices have edges with every other vertex of the same clique. In graph theory a clique is also present if there is one edge that connects two different vertices. But the cluster constructor tries to find the maximal clique that is the largest subgraph that fulfills the requirements of a clique. Hence, smaller cliques that are subsets of larger cliques are discarded. Although the term *clique* is usually reserved for undirected graphs, that should elucidate the problem from a mathematical point of view. In contrast to a classic clique problem, what differs in the case of the WGd clusters is that the underlying graphs are directed and a clique is accepted as such only when all vertices have inbound *and* outbound edges to every other member of the clique.

3. Discussion: advantages and limitations

The cross-reference clusters are a perfect match for the dictionary's layout. They reflect its workflow, which does not follow the alphabet, but focuses on topic domains, word fields, and word families. Against this background, it is no surprise that the rules, as stated in section 2, reveal a host of reciprocal reference structures even in the project's early stages. Therefore, the technical solution (i. e. the cross-reference clusters) is very much in line with the project design and with the microstructure of its entries.

We believe that the outcome also satisfies the prerequisites as formulated in section 1. The clusters' shape is clear and subtle; it guides viewers by highlighting and promoting those lemmas that are more pertinent than others in a given context. Especially their perspectivized form, when the lemmas are enriched with details about semantic relations, seems to have much value. But we can still only surmise that this is the case. A comprehensive study, which tracks the way users interact with the WGd website, would be highly desirable.

We have already hinted at some of the limitations to this approach: The structure of WGd entries, which is quite loose for a dictionary, imposes some technical issues. But as the clusters are clearly an addendum, whereas the text is the gist of an entry, we do not intend to change much in that respect. Although the analysis could gain a bit from a more in-depth markup of every reference's relational meaning, we will probably never force multi-word

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

entries to allot separate parts of the entry to one lemma alone. That would clearly contradict the project's mission and knowledge interest.

Of more concern, though, is that the cross-reference clusters don't scale well. It is hard to imagine a cluster with dozens of lemmas at its center. In that case, thresholds can reduce their size. We also are aware that the multitude of clusters that arise from this approach will gradually become overwhelming. Therefore, we already restricted analysis to topic boundaries. That means that, in practice, we don't take into account the whole reference network but subdivide it into thematic chunks. That does not exclude lemmas from, say, the topic domain *economy* from appearing in a cluster that was calculated for lemmas that pertain to *politics and society* (which is possible). But we do exclude lemmas from foreign topic domains during the detection of cluster centers.

This leads to another limitation. The calculation of cluster centers can be quite expensive in terms of computational workload — a well-known issue pertaining to the calculation of cliques. The current approach relies on analysis of all possible centers or cliques that can be derived from the references to a lemma. For example, there are currently five different lemmas that point to *elitär*: *Clique, Cliquenwirtschaft, Elite, Establishment,* and *Koterie.* In theory, every conceivable combination of these six lemmas (*elitär* has to be included) can form a cluster center. Therefore, possible combinations are as follows:

1) elitär, Clique, Cliquenwirtschaft, Elite, Establishment, Koterie
2) elitär, Clique, Cliquenwirtschaft, Elite, Establishment
3) elitär, Clique, Cliquenwirtschaft, Elite, Koterie

…

57) Establishment, Koterie

That should give an idea of what this actually means, as the number of combinations is already 57 in this example.[10] When we have to deal with six referring entries, the number rises

[10] There are 15 unique combinations with 2 lemmas, 20 with 3, 15 with 4, 6 with 5, and 1 with 6. A self-contained sample code that fills an array with all imaginable combinations looks as follows (in this case written in JavaScript):

```
let comb = [
  [
    [
     "elitär", "Clique", "Cliquenwirtschaft", "Elite", "Establishment", "Koterie",
    ],
  ],
];
let lemmas = 0;
let currentComb = [];
function makeComb (len, start) {
  if (len === 0) {
    comb[comb.length - 1].push([...currentComb]);
    return;
  }
  for (let i = start; i <= comb[0][0].length - len; i++) {
    currentComb[lemmas - len] = comb[0][0][i];
    makeComb(len - 1, i + 1);
  }
}
```


This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

to 120. One can easily imagine that this quickly leads to staggering figures and a huge computational workload. There are some tricks, though, to alleviate that problem. The 57 combinations in this example are not actually checked, as 53 of them can be discarded quickly. Consequently, this limitation does not pose a severe problem at the moment. But if one had to master much larger numbers of references and entries, it definitely would.

WGd is clearly not the first project that offers users visualizations with a navigational purpose. One might think of the word clouds in DWDS entries (*DWDS-Wortprofil*),[11] or the word graph in Wortschatz Leipzig.[12] At a first glance, these visualizations seem akin to the WGd clusters, as they are both an informational tool and a navigational device for users. But the underlying analyses are very different in terms of source data. Where WGd deals with manually set references by lexicographers, those projects operate with huge data sets and reveal cooccurrences.

On the contrary, the Semagrams of the Algemeen Nederlands Woordenboek (General Dutch Dictionary, ANW) have a lot to do with semantic relations. They are a good example of how a tight lexical structure can alleviate entries' findability. The ANW's Semagrams are highly structured lists of meanings that fill a predefined number of slots (Moerdijk et al. 2008). Each slot represents a semantic class and gives short informational descriptions like "size: is big", "place: is kept on a farm" for the entry *koe* ("cow").[13] Additionally, nonvisible data fields store keywords, synonyms and "relevant words" (ibid., p. 20). These are used for advanced database queries. But the ANW's Semagrams stem from a completely different lexicographical approach and are therefore not applicable to WGd.

A bit more in line with the style of WGd is *elexiko*.[14] The dictionary also deploys different entry types, including word group articles (*Wortgruppenartikel*) and multi-word entries that deal with sense-related lemmas (*sinnrelationale Paare und Gruppen*). Some of those entries include raster graphics that visualize semantic relations in the form of complex stemmas or intersected word fields. However, it seems that these graphics have not been calculated, but individually created for the specific purpose of an article in question. Contrary to WGd, *elexiko* uses a tighter XML structure for the cross-reference markup (cf. Meyer/Müller-Spitzer 2010). Therefore, an automatic visualization of the semantic relations should not pose much difficulty. With the exception of some similarities to the WGd macrostructure, the standard entries of *elexiko* rely much less on continuous text. Instead, they offer an astounding multitude of word information. In that sense, the dictionary is much more in line with the ANW.

In short: None of these projects drew on the same solution. That is certainly not a stunning observation, as all these tools were devised on a different lexicographical basis and out of diverging knowledge interests. Eventually, it seems that every project needs a solution of its own that matches both its scientific approach and its dictionary's microstructure.

```
for (let i = comb[0][0].length - 1; i >= 2; i--) {
  comb.push([]);
  lemmas = i;
  currentComb = [];
  makeComb(i, 0);
}
```

[11] https://www.dwds.de/d/ressources#wortprofil.

[12] https://wortschatz.uni-leipzig.de/en.

[13] https://anw.ivdnt.org/article/koe.

[14] https://www.owid.de/docs/elex/start.jsp.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

References

Blensenius, K. et al. (2021): Finding gaps in semantic descriptions: Visualisation of the cross-reference Network in a Swedish monolingual dictionary. In: Proceedings of the eLex 2021 Conference, Brno, pp. 247–258. https://elex.link/elex2021/proceedings-download/ (last access: 25-03-2022).

Meyer, P./Müller-Spitzer, C. (2010): Consistency of sense relations in a lexicographic context. In: Proceedings of the Workshop "Semantic Relations. Theory and Applications". Malta, pp. 37–46. http://www.lrec-conf.org/proceedings/lrec2010/workshops/W9.pdf (last access: 24-03-2022).

Moerdijk, F. et al. (2008): Accessing the ANW dictionary. In: Coling 2008: Proceedings of the workshop on Cogintive Aspects of the Lexicon (CogALex 2008), Manchester, pp. 18–24.

Therón, R. et al. (2014): Highly interactive and natural user interfaces: enabling visual analysis in historical lexicography. In: Proceedings of DATeCH 2014. Madrid, pp. 153–158. https://dl.acm.org/doi/proceedings/10.1145/2595188 (last access: 24-03-2022).

Contact information

Nico Dorn
Akademie der Wissenschaften zu Göttingen
ndorn@gwdg.de

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Mireille Ducassé/Archil Elizbarashvili

FINDING LEMMAS IN AGGLUTINATIVE AND INFLECTIONAL LANGUAGE DICTIONARIES WITH LOGICAL INFORMATION SYSTEMS

The case of Georgian verbs

Abstract Looking up for an unknown word is the most frequent use of a dictionary. For languages both agglutinative and inflectional, such as Georgian, this can be quite challenging because an inflected form can be very far from the lemmas used by the target dictionary. In addition, there is no consensus among Georgian lexicographers on which lemmas represent a verb in dictionaries. It further complicates dictionaries access. Kartu-Verbs is a base of inflected forms of Georgian verbs accessible by a logical information system. It currently contains more than 5 million inflected forms related to more than 16,000 verbs for 11 tenses; each form can have 11 properties; there are more than 80 million links in the base. This demonstration shows how, from any inflected form, we can find the relevant lemma to access any dictionary. Kartu-Verbs can thus be used as a front-end to any Georgian dictionary.

Keywords E-dictionary; lemma; Georgian language; under-resourced language; inflected forms; logical information systems; semantic web

1. Introduction

The survey reported in Kosem et al. (2019), involving 10,000 people of 26 countries, shows that looking up for an unknown word is the most frequent use of a monolingual dictionary. For languages both agglutinative and inflectional, such as Georgian, this can be quite challenging because an inflected form can be very far from the lemmas used by the target dictionary. As discussed in Ducassé (2020), conjugation can modify any part of the form of Georgian verbs and to go from a conjugated form to a lemma requires a good knowledge of Georgian grammar. Moreover, the lemmatization of verbs in Georgian dictionaries is still an open problem, see Margalitadze (2020) and Gippert (2016). Most dictionaries use "Georgian infinitives"[1] as entries. The "Comprehensive Georgian-English Dictionary" (Rayfield et al. 2006) presents, for all verbs, the infinitive as well as the 3rd person singular in the present and the future. The Georgian-German dictionary of Tschenkéli et al. (1965) uses the abstract verb root under which all sub-paradigms are listed. Thus, depending on the target dictionary, starting from inflected form "vikiraveb" (ვიქირავებ, «I will rent»), a user would have to find Georgian infinitive "kiraoba" (ქირაობა), 3rd person present "kiraobs" (ქირაობს), 3rd person future "ikiravebs" (იქირავებს), or root "kirav" (ქირავ). Note that this verb is a relatively "easy" one.

Kartu-Verbs is a knowledge base of inflected forms of Georgian verbs[2] integrated within a semantic web tool, Sparklis, a platform to easily implement logical information systems.

[1] What we call "Georgian infinitive" is, strictly speaking, a verbal noun. This is the form that comes closest to an infinitive. We use this term because it is easier to understand for non-linguist English-speaking target users.

[2] Kartu-Verbs can be accessed at https://www-semlis.irisa.fr/software/georgian-verb-inflected-forms-base/.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Logical information systems allow users to retrieve information on the facets of their choice by progressively refining their queries using suggestions (Ferré 2017). The base can be easily navigated in all directions: from an inflected form to an infinitive, and conversely from an infinitive to any inflected form; from components to forms and from a form to its components. Thanks to our collaboration with Paul Meurer, we are in the process of integrating the very rich knowledge about Georgian verbs underneath the Georgian part of his INESS::XLE-Web platform (Meurer 2007).[3] The integration is not yet completed. Kartu-Verbs contains, however, already more than 5 million inflected forms related to more than 16.000 verbs for 11 tenses. Each form can have 11 properties. There are more than 80 million links in the base. All keywords can be displayed both in Georgian and Latin characters. This version of Kartu-Verbs is almost 3 orders of magnitude larger than the previous version that only contained thousands of inflected forms related to hundreds of verbs. Ducassé (2020) mainly illustrates how to use Kartu-Verbs to gain knowledge about Georgian conjugation. This demonstration shows in details how, from any inflected form, one can find the relevant lemma(s) to access any dictionary. Kartu-Verbs can thus be used to build a front-end to any Georgian dictionaries.

2. Finding relevant entries from inflected forms

This section explains how to use Kartu-Verbs to find in a few clicks 4 different lemmas for 4 different dictionary organizations, starting from an inflected form. In addition, we show how to get verb variants from a given root. We use a Latin transliteration following Georgian government recommendations.[4] The base can also be searched using the Georgian alphabet.

Figure 1 shows the 3 areas of the Kartu-Verbs user interface, from left to right and top to bottom: the query, the suggestions and the results. A basic query is displayed to find 4 of the features of inflected forms: person, number tense and (surface) form. The "Suggestions" area, is itself divided into sub-areas, only 2 are of interest to us for this demonstration. On the left, the "Types and Relationships" sub-area offers features that can still be added to the query; on the right, the "Identities or values" box suggests some of the inflected form identifiers that match the query. Suppose the user is looking for lemmas for "vikiraveb" ("ვიქირავებ"). We can see in the figure that, by entering "vikiraveb_" in the search part of the "Identities or Values of the thing" sub-area, two forms are accessible for two different tenses, their identifiers are "vikiraveb_future_1sg_kiraoba" and "vikiraveb_present_1sg_dakiraveba". Note that the results area has already adjusted to the suggested values. Sometimes this is enough to find the information we are looking for. As discussed above, depending on the targeted dictionaries, the Georgian infinitive, the 3rd-person singular present or future, as well as the root, would be needed. In the relations sub-area of suggestions, on the left, we see that the "Georgian infinitive" and the "root" features are accessible.

[3] INESS::XLE-Web is a powerful tool dedicated to linguists. It is able to parse sentences and produce syntactic trees for a number of languages, including Georgian. It contains a lot of interesting information but presented in a form not really accessible to beginners.

[4] http://www.enadep.gov.ge/uploads/Bulletin_II_2019-2020.pdf.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

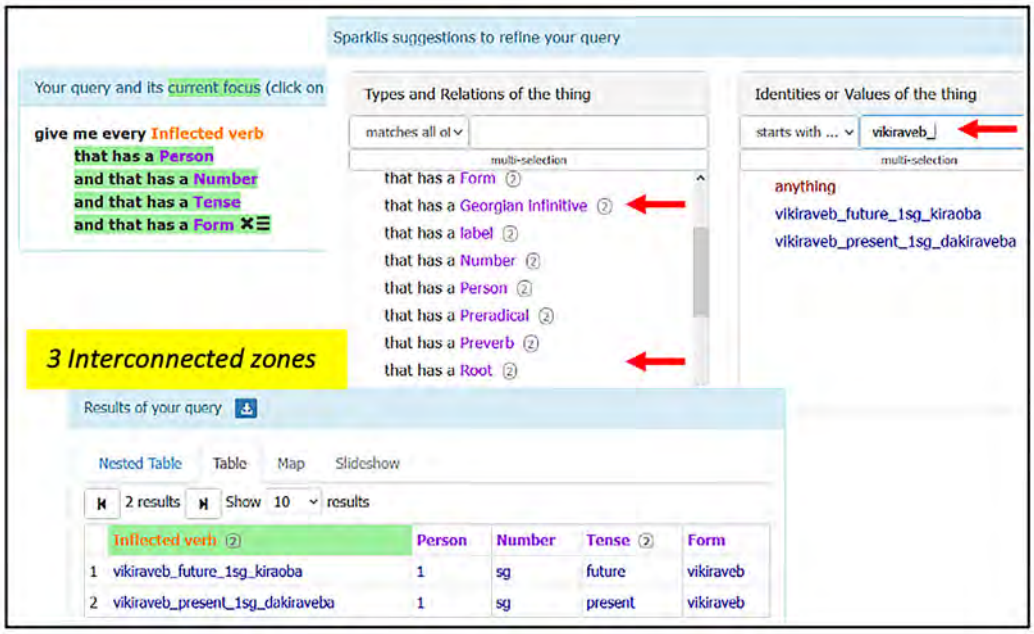


Fig. 1: Searching for inflected form "vikiraveb"

Figure 2 shows the query and result areas after the user clicked on the 3 suggestions. The query has been automatically updated. We see that the inflected form is "vikiraveb". The "Georgian infinitive" and "root" features have been added. In the results area, two columns have appeared giving the two pieces of information sought. The Georgian infinitive is "kiraoba" (ქირაობა) for the present form and "dakiraveba" for the future form. The root is "kirav" for both forms. We can also see that both forms correspond to the first-person singular. Clicking on the arrow at the right of each of the Georgian Infinitive opens a new window to access the translate.ge site. We can see that "kiraoba" can mean "to hire" or "to rent".

Fig. 2: Two lemmas for inflected form "vikiraveb" (Georgian infinitive and root)

We have thus found in a few clicks two of the four lemmas. Figure 3 displays a refined query to find the two other lemmas for verb "kiraoba", its 3rd-person singular forms present and future. From the previous query, we have kept the Georgian infinitive and the number by clicking on "kiraoba" and "singular" in the table of results. We specified (by clicking in the suggestions) that we wanted third-person forms. We took advantage of the logical capabilities of Sparklis queries to ask for tense to be either present or future (using the ≡ menu

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

icone next to query elements). We have also asked to see 5 features in the result table: Root, Form, Person, Number and Tense. The results area displays 3 form identifiers because the verb has two different forms at present. Note that if we had not specified the tense, we would have gotten a conjugation table with the third person singular forms of the verb "kiraoba" at all tenses. We can see on Figure 3 that the same verb can have different roots at different tenses ("kira" at present or "kirav" at future.) We could investigate further which root is used for which tenses by clicking on one root not specifying the tense.

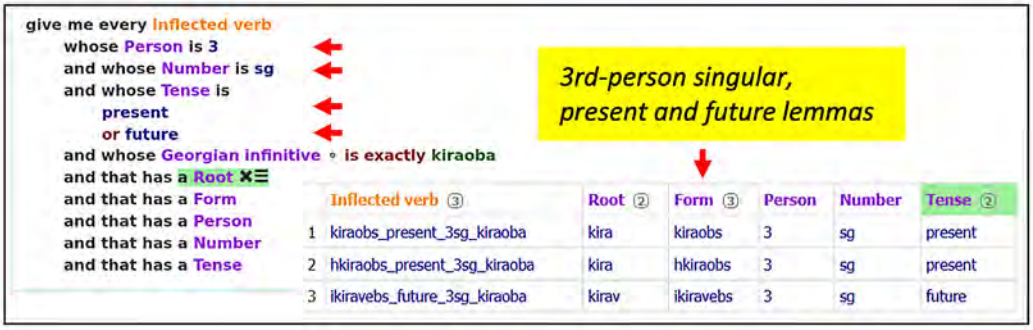


Fig. 3: Two more types of lemmas for vikiraveb

In Georgian, many verb variants can be constructed with a given root. Figure 4 shows a query to find all the 3rd person singular future of verbs whose root is either "kira" or "kirav". There are 15 different forms with different preverbs and preradicals. This shows the importance of morphological information for agglutinative languages. We could click on any of the morphemes to refine the query.

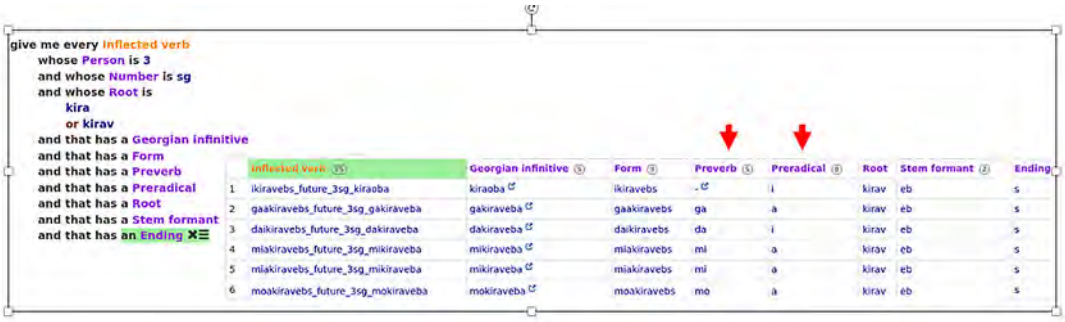


Fig. 4: Finding all the verbs with given roots

3. Discussion

In this demonstration, even for a relatively simple verb we have encountered a number of "homonyms", namely entries with identical surface forms with different attributes (for example, "vikiraveb" can be either at present or future). Figure 1 illustrates how the form identifiers help to distinguish homonyms (for example "vikiraveb_future_1sg_kiraoba" vs "vikiraveb_present_1sg_dakiraveba"). There can also be "synonyms", namely entries sharing their properties with different surface forms (for example "kiraobs" and "hkiraobs" are

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

two different surface forms for verb "kiraoba" at 3rd singular present). In Figures 3 and 4 we can see that synonyms can easily be captured thanks to the query mechanism that can traverse the base in all directions and easily go from components to entries.

The queries can use any combination of logical operators "AND", "OR", "NOT" at any place as the user can specify the focus (scope) of an operator. There are also aggregate operators, not demonstrated here. We have illustrated how to add features in the query. The user can also easily remove features that are not of interest to him at a given moment by clicking on the "x" attached to them as can be seen next to "tense" in Figure 3. As illustrated above, all queries are constructed using suggestions. Users have nothing to invent. They can use filters to help Kartu-Verbs come up with relevant suggestions, then queries are built only by clicking on suggestions that are necessarily relevant. The benefits are three-fold, first, it is easier to find something in a list than typing it, second, users cannot make typos, and finally, as a direct result, queries can never yield an empty result. This is a very strong property due to the powerful mechanisms of Sparklis.

Some linguists provide comprehensive tables of inflected forms, for example the "Georgische Verbtabellen" by Chotiwari-Jünger et al. (2010) or the books of the "Biliki" series by Nana Shavtvaladze[5]. The latter offer conjugation tables of several types as an appendix to the lessons. They contain invaluable information. However, learners have to go through different books to find relevant information. When searching for an inflected form, learners must check each of the over 10,000 entries. Furthermore, the inflected forms use the Georgian alphabet, which is a big hurdle for beginners. Exceptions, quite common, cannot always be anticipated from the sample tables. Unlike tables, in Kartu-Verbs there are no predefined usages. Any field can be used to search the database at any moment and the whole information is accessible from the query.

Other systems provide all inflected forms for inflectional and agglutinating languages. For example, in Verbel, the inflection dictionary for Polish, morphological structure and inflectional properties are used to structure verb forms into "flexemes" (Czerepowicka 2021). We still have to investigate if this notion of flexeme could help us integrate more data with reasonable performances. In DeCesaris (2021), the author advocates to add more morphological information in monolingual dictionaries, in particular because users may not have enough knowledge about inflectional rules and derived words may have acquired additional nuances of meaning. Morphological information is also considered crucial by de Schryver/Chishman/DaSilva (2019) for languages with complex morphology such as Bantu. In Kartu-Verbs, not only do we provide morphological information, but each morpheme, like any piece of information, is an URI that can be used to lead to another piece of information.

A dedicated front-end to dictionaries is under development so that users do not have to see the queries for the usual searches, while being able to access sophisticated queries when needed. We also plan to integrate more features, in particular from the INESS::XLE-Web platform (Meurer 2007). It raises issues both in terms of performances and data structures.

4. Conclusion

In this article, we have illustrated how versatile and powerful LIS navigation mechanisms are. We have also shown how these mechanisms can help users of Kartu-Verbs easily obtain

[5] Biliki, Georgian Language For English Speakers. See http://lsgeorgia.com.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

information about the verbs they encounter in Georgian texts whatever their form. In particular, finding the relevant lemma(s) for an entry in a given dictionary is no longer a problem. Kartu-Verbs can thus be used as an interface for any Georgian dictionary, regardless of the lemmatization principles the dictionary uses for verbs.

References

Chotiwari-Jünger, S. et al. (2010): Georgische Verbtabellen. Hamburg.

Czerepowicka, M. (2021): The structure of a dictionary entry and grammatical properties of multiword units. In: Electronic lexicography in the 21st century: eLex 2021, pp. 200–215.

de Schryver, G.-M./Chishman, R./DaSilva, B. (2019): An overview of digital lexicography and directions for its future: An interview with Gilles-Maurice de Schryver. Universidad do Vale de Rio dos Sinos.

Ducassé, M. (2020): Kartu-verbs: A semantic web base of inflected Georgian verb forms to bypass Georgian verb lemmatization issues. In: Gavriilidou, Z./Mitsiaki, M./Fliatouras, A. (eds.): Proceedings of XIX EURALEX International Congress, Volume 1, pp. 81–89.

Ferré, S. (2017): Sparklis: An expressive query builder for SPARQL end-points with guidance in natural language. In: Semantic Web: Interoperability, Usability, Applicability 8 (3).

Gippert, J. (2016): Complex morphology and its impact on lexicology: the Kartvelian case. In: Margalitadze, T./Meladze, G. (eds.): Proceedings of the 17th EURALEX International Congress, Tbilisi, Georgia, pp. 16–36. Ivane Javakhishvili Tbilisi University Press.

Kosem, I./Lew, R./Müller-Spitzer, C./Ribeiro Silveira, M./Wolfer, S./Dorn, A./Gurrutxaga, A./Ceberio, K./Etxeberria, E./Lefer, M.-A. et al. (2019): The image of the monolingual dictionary across Europe. In: International Journal of Lexicography 32 (1), pp. 92–114.

Margalitadze, T. (2020): Lexicography of Georgian: a brief overview. Language@Leeds Working Papers in Linguistics, University of Leeds.

Meurer, P. (2007): A computational grammar for Georgian. In: Logic, language, and computation. 6th International Tbilisi Symposium on Logic, Language, and Computation. Berlin, pp. 1–15.

Rayfield, D./Apridonze, S./Chanturia, A./Amirejibi, R./Broers, L./Chkhaidze, L./Margalitadze, T. (2006): A Comprehensive Georgian-English Dictionary. London.

Tschenkéli, K./Marchev, Y./Flury, L. (1965): Georgisch-deutsches Wörterbuch, Volume 2. Zürich.

Contact information

Mireille Ducassé
IRISA-INSA Rennes
mireille.ducasse@irisa.fr

Archil Elizbarashvili
Ivane Javarishvili Tbilisi State University
archil.elizbarashvili@tsu.ge

Acknowledgements

We are indebted to Paul Meurer who let us transfer the invaluable data from his private Clarino base. We thank Mikheil Sulikashvili for his technical support to transfer Clarino information into Csv files. We warmly thank Sébastien Ferré for his strong support to use Sparklis.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Velibor Ilić/Lenka Bajčetić/Snežana Petrović/Ana Španović

SCyDia – OCR FOR SERBIAN CYRILLIC WITH DIACRITICS

Abstract In the currently ongoing process of retro-digitization of Serbian dialectal dictionaries, the biggest obstacle is the lack of machine-readable versions of paper editions. Therefore, one essential step is needed before venturing into the dictionary-making process in the digital environment – OCRing the pages with the highest possible accuracy. Successful retro-digitization of Serbian dialectal dictionaries, currently in progress, has shown a dire need for one basic yet necessary step, lacking until now – OCRing the pages with the highest possible accuracy. OCR processing is not a new technology, as many open-source and commercial software solutions can reliably convert scanned images of paper documents into digital documents. Available software solutions are usually efficient enough to process scanned contracts, invoices, financial statements, newspapers, and books. In cases where it is necessary to process documents that contain accented text and precisely extract each character with diacritics, such software solutions are not efficient enough. This paper presents the OCR software called "SCyDia", developed to overcome this issue. We demonstrate the organizational structure of the OCR software "SCyDia" and the first results. The "SCyDia" is a web-based software solution that relies on the open-source software "Tesseract" in the background. "SCyDia" also contains a module for semi-automatic text correction. We have already processed over 15,000 pages, 13 dialectal dictionaries, and five dialectal monographs. At this point in our project, we have analyzed the accuracy of the "SCyDia" by processing 13 dialectal dictionaries. The results were analyzed manually by an expert who examined a number of randomly selected pages from each dictionary. The preliminary results show great promise, spanning from 97.19% to 99.87%.

Keywords OCR; Cyrillic; Serbian language; retro-digitization; convolutional neural networks

1. Introduction

In the Institute for the Serbian language of SASA, several lexicographic projects – descriptive, etymological, historical, dialectal, neological, etc. – are currently ongoing and still compiled in the traditional way. The lexical material they are based upon includes numerous dictionaries and scientific monographs, which have to be consulted in the paper edition. The vast majority of these dictionaries and monographs (tens of thousands of pages), dedicated to compiling and analyzing dialectal lexis, and describing dialectal features, are written in Cyrillic, containing accents, diacritics, and other non-standard characters. We should bear in mind that the Serbian language is in the position of being low-resourced in the field of digital infrastructure and digitized language resources (for example, in the Institute, no dictionary is corpus-based nor corpus-driven, and no tools for writing or editing dictionaries in the digital environment are used, etc.). Even though some serious first steps have been taken towards applying new technologies to our lexicographic legacy[1] and into the dictionary-making process,[2] we were well aware that this obsolete methodology may question the relevance of research results and downgrade the scientific level of publications. Therefore,

[1] See dictionary platforms Raskovnik and Prepis.

[2] Certain significant steps have been taken also towards digitization of the *Dictionary of the Serbo-Croatian Standard and Vernacular Language of the Serbian Academy of Sciences and Arts* (Stijović/Stanković 2018). Some volumes passed the OCR processing, and manual correction afterwards. However, there is no data on OCR output precision, or how many working hours were spent on corrections (Stanković et al. 2018, p. 942).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

we decided to take a broader approach to improve our work – to retro-digitize this vast number of scientific dictionaries and monograph studies of fundamental importance for lexicographic work. That will enable us to create a multifunctional lexicographic database and different corpora and use dialectal material to produce various dictionaries, scientific papers, etc. One of the significant accomplishments of this process of retro-digitization, in the long run, should also be the promotion of dialects and vernaculars, especially in modern-day society. However, the biggest obstacle when attempting to retro-digitize Serbian dialectal dictionaries was the lack of machine-readable versions of paper editions, implying that we needed to complete one essential step before venturing into the dictionary-making process in the digital environment – OCRing the pages with the highest possible accuracy.

Optical Character Recognition (OCR) is a process that allows data extraction from a scanned document or image file. In this process, the printed or handwritten text on the scanned document is converted to a machine-readable format. OCR processing is not a new technology, and there are many open-source and commercial software solutions that can reliably convert scanned images of paper documents into digital documents. Even so, available software solutions are usually efficient enough to process scanned contracts, invoices, financial statements, newspapers, and books. In cases where it is necessary to process documents containing accented text and precisely extract each character with diacritics, such as dialectal dictionaries written with Cyrillic letters, such software solutions are not efficient enough.

1.1 Why OCR?

Although double-keying is the most accurate way for transcription, it is very time-consuming and – in the case of dialectal and historical dictionaries, with text too complex for non-experts – costly because it requires additional corrections, usually more than one. This is based on our previous work experiences digitizing five dialectal dictionaries currently available on *Raskovnik*. Therefore, to overcome this problem, we decided to invest in developing an OCR software called "SCyDia" – *Serbian Cyrillic with Diacritics*. By now, we ran the "SCyDia" software on 14 dictionaries and monographs with more than 15,000 pages combined, but we intend to use it on hundreds of thousands of pages more.

Since the accuracy of OCR varies from 97,19% to 99,87%, some dictionaries would be reasonably quick to verify manually. On the other hand, the worst result of a 2,81% error rate in one dictionary means that a page of 3000 characters has 84,3 errors which can be time-consuming and too expensive to correct. We have opted for a less-than-perfect gradual approach in these cases by correcting only the headword lemmas[3] in the first phase. In this way, we could make our database "searchable" while still keeping the cost reasonably low.

1.2 Related Work

Klyshinsky/Karpi/Bondarenko (2020) compares neural network software used to restore diacritics in six languages such as Croatian, Slovak, Romanian, French, German, Latvian, and

[3] The objective to have a fully and precisely corrected version of the digitized material in Cyrillic with diacritics and other non-standard characters prior to start using it in a lexicographic work process is utmost time-consuming and unrealistic from the financial perspective. See for example Vitas/Krstev (2015, p. 109).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Turkish. The recognition accuracy usually ranges from 95 to 99%, depending on the letter; some letters have relatively low accuracy.

Hussain et al. (2014) present the results of using the Tesseract engine for OCR processing of pages written by Urdu Nastalique (a very complex and cursive writing style of Arabic script); without any modifications, the Tesseract achieves an accuracy of 66%, and with additional modifications, the accuracy is increased to 97%.

Cristea et al. (2020) present the results of a solution based on several types of neural networks (such as The Region Proposal Network (RPN), ResNet, Faster R-CNN) for OCR processing of old Romanian documents written in Cyrillic.

Rijhwani/Anastasopoulos/Neubig (2020) describes post-correction methods where the goal is to reduce the number of errors that occur during OCR processing that most often happen due to low-quality scanning, physical deterioration of paper book, or different styles of font.

In their research, Krstev/Stanković/Vitas (2018) present the process of restoring diacritics in Serbian texts written in degraded Latin script, and the presented solution relies on the comprehensive lexical resources for Serbian: the morphological electronic dictionaries, the Corpus of Contemporary Serbian and local grammars.

In their research, O'Brien/Haddej (2012) present a project where the functionality of OCRopus software has been expanded to support the recognition of mathematical symbols and unique linguistic alphabets (e.g., Hungarian letters) while the extended version supports UTF-8 character encoding. The accuracy of the original version trained only with English characters was 86%; in the extended version, the accuracy increased to 93,5%.

1.3 An overview of the "SCyDia" software

This paper will present the OCR software "SCyDia", a web-based software solution that relies on open-source software Tesseract V5 in the background. The software is developed to overcome the problem of not having OCR software efficient enough to process documents containing accented text and precisely extract each character with diacritics. Finally, we will demonstrate the organizational structure of the software and the first results.

The paper is organized as follows. Section 2 contains implementation details, details about used convolutional neural networks (CNN) and datasets, and a description of modules for semi-automatic text correction. After that, section 3 presents the results. Further plans are presented in section 4. Finally, the last section contains conclusions.

2. Implementation of SCyDia

The "SCyDia" OCR software is developed as a web application; an overview of the algorithm is presented in Figure 1. It allows the user to see a list of scanned pages and select pages for OCR pressing or text correction (proofreading).

The web application (1) allows the user to choose which scanned pages will be processed. The selected images of the scanned text pages (2) are forwarded to the Python application. OCR processing in the initial step uses Tesseract (3), which generates a text file (6) with recognized text without diacritic signs. Tesseract also returns coordinates of bounding boxes around individual letters. The coordinates of bounding boxes are usually concretely determined. Occasionally, instead of one letter inside the bounding box, it may contain two,

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

three, or even more letters; sometimes, the bounding box can contain halves of two adjacent letters.

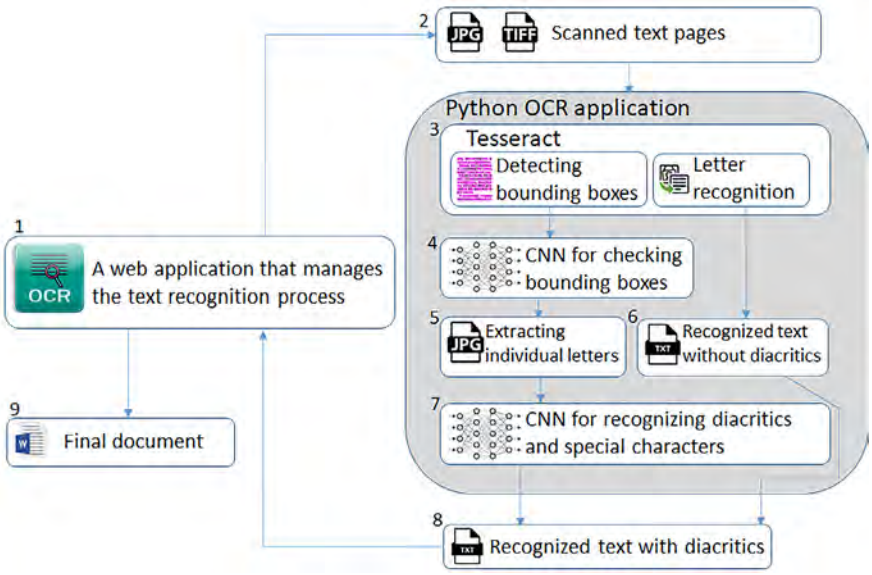


Fig. 1: Overview of ScyDia software

The convolutional neural network (4) can check whether the bounding box contains only one letter as expected, and if there is more than one, it returns information on how many letters are inside the bounding box. For example, detected bounding boxes with more than one letter are divided into an appropriate number of smaller bounding boxes containing one letter.

In Figure 2, the correctly determined bounding boxes with one letter are shown in blue. Those boxes that initially contained two letters and were divided into two parts are shown in green, and boxes with three letters are divided into smaller boxes are shown in yellow color. Bounding Border boxes where multiple letters are detected are automatically divided into the appropriate number of parts to contain one letter using the Python script.

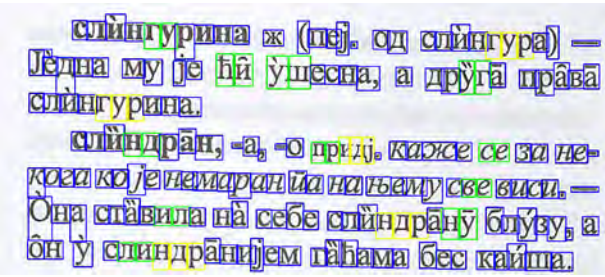


Fig. 2: Detected bounding boxes around letters

In Figure 1, Python script (5) uses bounding boxes coordinates to extract individual letters' images. The convolutional network (7) processes those images of individual letters and tries to detect whether they contain diacritic signs. Also, this network can be used to detect letters that Tesseract has difficulty recognizing correctly, such as italic letters *ī ū ш в*. In the final step, the Python function tries to match each letter from a text file with the information

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

provided by the convolutional network when processing extracted images of those letters. The result of that function represents a new text file containing letters with diacritic signs. For example, "SCyDia" software generates text in format UTF-8 plaintext; letters with diacritics consist of two characters, one character for the letter and the other for the diacritical character (symbol).

2.1 Network Configuration and Datasets

The "SCyDia" OCR application uses two convolutional neural networks, CNN for checking bounding boxes and CNN for detecting diacritics. These two networks have similar configurations, and they differ in the number of outputs.

The **CNN used for detecting diacritics** takes a 48×32×1 matrix as input; it contains three convolutional layers. The first layer contains 16, the second 32, and the third layer contains 64 3×3 kernels with *ReLu* activation. After each layer, a max-pooling layer with a pooling size of 2×2, a dropout probability of 0.25 is placed. Three fully connected layers follow these convolutional layers: the first layer contains 128 nodes, the second 64 nodes, and the third layer contains 32 output neurons. After each layer is placed, the dropout layer with a dropout probability of 0.25. Finally, the output layer contains 30 nodes, Figure 3. The values obtained at the network output have the following meaning: the first value indicates whether the letter contains diacritic signs, the second whether the letter is correct (sometimes the bounding box is not placed correctly around the letter), and the following 15 values detect the type of diacritic signs, the remaining values are used to detect letters Tesseract does not recognize correctly, for example, letters (ѣәѣŋʒ, and italic letters such as *ī ū ѿ в*).

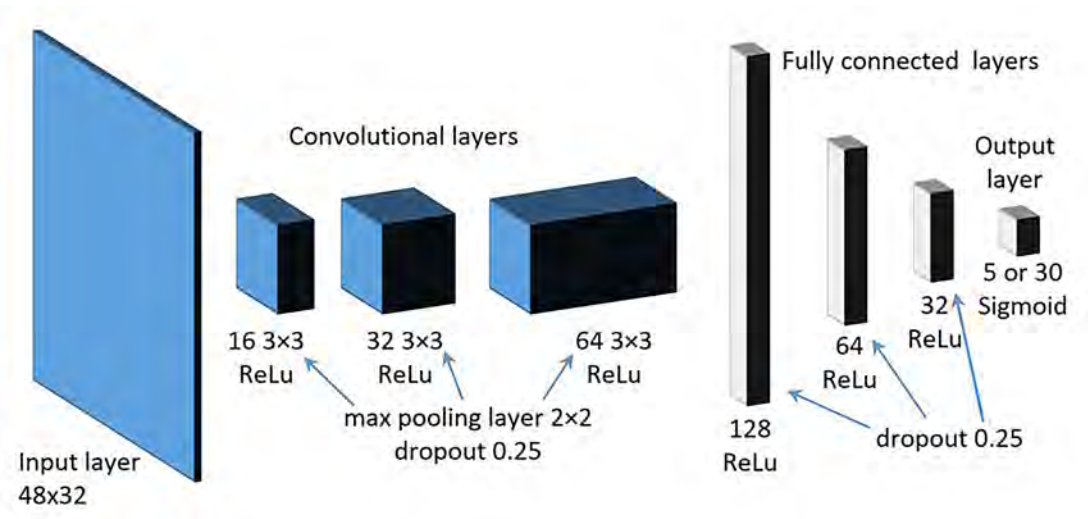


Fig. 3: Structure of convolutional networks

Datasets for CNN used for detecting diacritics are generated by collecting cropped individual letters from scanned pages. This dataset contains:[4]

[4] It's worthwhile noting that all scholar dictionaries in Serbian, and even most of the popular ones, are using characters with diacritics.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Group of Cyrillic letters:

- Standard set of Cyrillic letters,
- Letters that have diacritics above the letters, for example:

  ȁ á ȃ â ā ȧ ă ã ấ ǎ ȧ
- Letters that have diacritics below the letters, for example:

  ạ ậ ặ ḽ
- Letters that have diacritics above and below the letters,
- Cyrillic letters that do not belong to the standard set of symbols that Tesseract cannot recognize, for example:

  ьәҽŋʒ Tesseract incorrectly recognizes these letters as: Б о ҂ о з
- Letters where one letter consists of two characters, for example: д͡з

The **CNN used for checking bounding boxes** has a similar configuration; the output layer of that network contains 5 nodes, Figure 3. The values obtained at the network output have the following meaning – the first value indicates that bounding box is around one letter, and the second value indicates that bounding box is around two letters. The third value indicates that the bounding box is around three letters, the fourth value indicates more than three letters, and the fifth value is used to detect invalid letters; for example, there are two halves of consecutive letters within the boundary frame.

Datasets for CNN used for checking bounding boxes are also generated by collecting cropped letters from scanned pages. This dataset contains examples of how an adequately extracted letter looks, examples of when two or three letters are extracted together, and examples of images with incorrectly extracted letters when two halves of a letter are in a boundary field.

Adam optimizer is used for both networks. The duration of training was limited to 50 epochs, with two additional parameters: *ReduceLROnPlateau* with patience 10 and *EarlyStopping* with patience 25. Parameter *ReduceLROnPlateau* would reduce the learning rate if there were no improvement in the accuracy of the validation dataset for 10 epochs. *EarlyStopping* interrupts training if there is no improvement in the accuracy of the validation dataset for 25 epochs.

2.2 Manual and semi-automatic text correction (proofreading)

The primary purpose of the "SCyDia" application is OCR processing; besides that web application also provides a module for text correction (proofreading). That module allows manual and semi-automatic text correction. The window for **manual text correction** is divided into three fields (Fig. 4), the first field contains the recognized text, and it is an editable field; the second field contains cut-out images of paragraphs; in the third field, there is a complete picture of the scanned page on which the letters containing diacritics are marked.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

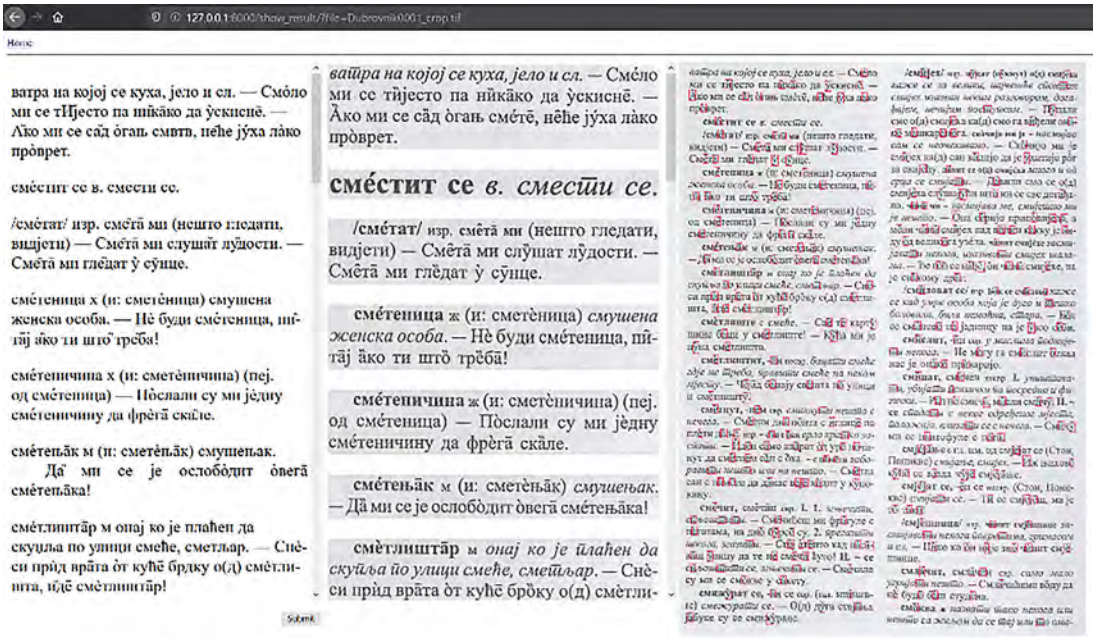


Fig. 4: Window for manual text correction

In order to **achieve semi-automatic text correction** (Fig. 5), the "SCyDia" application repeats OCR processing (3) of one page several times to create additional copies of text files that can be compared with each other. The algorithm for semi-automatic text correction starts by creating additional two copies (2) of the scanned page (1), the first image is rotated to the left by half a degree, and the second copy is rotated to the right half a degree. If they visually compare those images, humans will not notice the differences between the original scanned page and copies of that image rotated by half a degree. However, for OCR software, such a small difference causes misrecognized letters to appear in different places in the recognized text.

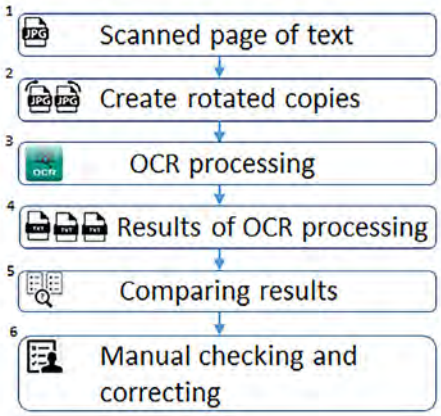


Fig. 5: Algorithm for semi-automatic text correction

In the next step, those text files (4) are compared with each other (5), and each detected difference is presented on the window for manual checking and correcting (6). In most cases, users can click on the button with the correct version of a word, Figure 6.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

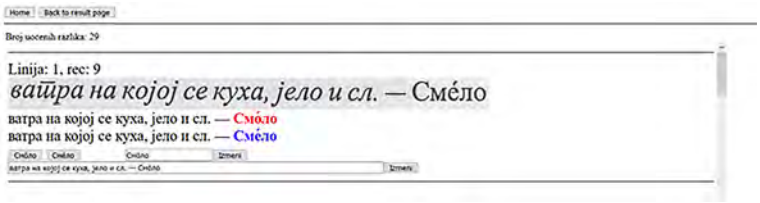


Fig. 6: User interface with results of semi-automatic text correction

The user interface with the results of the semi-automatic text correction contains following elements:

- the part of the scanned image with the text line where the difference is noticed,
- the text line where the difference is noticed from the original scanned page (word where the difference is presented in red),
- the text line where the difference is noticed from the rotated copy (word where the difference is presented in red),
- Button with a version of word from the first file,
- Button with a version of word from the second file,
- A text box that allows the user to manually correct an error if neither of these two versions is correct.

2.3 Usage of "SCyDia"

The "SCyDia" application has so far been used for processing over 15 000 pages of dialectical dictionaries of Serbian. The OCR process is conducted on a PC with Intel I9 12-core processor, with NVidia GeForce RTX 2070 SUPER graphic card. The "SCyDia" application can process eight pages in parallel, and each page is analyzed three times: first in its original shape and then skewed for half a degree left and right. On average, each page takes about half an hour to process. After the first batch of 14 dictionaries was processed, the results were analyzed. We have compiled a list of the most common problems for each dictionary. A list of letters and diacritics signs has been compiled, with the most common problems in each dictionary. Based on this list, an additional set of images with letters and diacritics will be generated to expand the dataset for training CNN used to detect diacritics.

3. Results

3.1 General characteristics of processed dictionaries

Table 1 provides an overall description of 13 dictionaries processed in the "SCyDia" application by showing some of their main characteristics relevant for the OCR, such as the possession of characters with diacritics in the headword, characters with diacritics in the citation, characters in italic, abbreviations, as well as characters in superscript.

The accuracy of OCR processing is evaluated by comparing the text generated by the OCR software with the reference text (manually typed text); the comparison is performed using a script, and the results obtained are shown in the following table.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

DICTIONARIES	CHARACTERS WITH DIACRITICS IN THE HEADWORD	CHARACTERS WITH DIACRITICS IN CITATION	CHARACTERS IN CURSIVE	ABBREVIATIONS	SUPERSCRIPT
Bašanović-Čečović (2010)	+	+	+	+	+
Boričić Tivranski (2002)	+	-	+	+	-
Bukumirić (2012)	+	+	+	+	-
Cvetanović (2013)	+	+	-	+	-
Cvijetić (2014)	+	+	+	+	-
Dalmacija (2004)	+	+	+	+	+
Dalmacija (2017)	+	+	+	+	+
Đoković (2010)	+	-	-	+	-
Rajković Koželjac (2014)	+	+	+	+	-
Ristić (2010)	+	+	+	+	+
RSGV (2000–)	+	+	+	+	-
Stanić (1990–1991)	+	+	+	+	+
Zlatković (2014)	+	+	+	+	-

Table 1: Overall description of dictionaries' complexity

As expected, characters with diacritics in the headword are present in each of the 13 dictionaries. Characters with diacritics in the citation are documented in most dictionaries (11 out of 13), except in Boričić Tivranski (2002) and Đoković (2010). 11 out of 13 dictionaries have characters in cursive, except Cvetanović (2013) and Đoković (2010). Abbreviations, such as grammatical ones, and locations and sources are present in all 13 dictionaries. Finally, superscript is found in 5 out of 13 dictionaries and missing from Boričić Tivranski (2002), Bukumirić (2012), Cvetanović (2013), Cvijetić (2014), Đoković (2010), Rajković Koželjac (2014), RSGV (2000–), and Zlatković (2014).

3.2 OCR processing accuracy

The accuracy of OCR processing was evaluated manually by experts. Although the "SCyDia" software provides semi-automatic detection of errors by comparing the slightly rotated versions to the original, we have decided to evaluate manually to ensure that the evaluation results are as accurate as possible. Semi-automatic error detection is beneficial for manual correction, but we cannot be sure that all errors are detected in this way. The experts counted all errors on the page and errors in "special" characters: letters with diacritics, italic, and specific abbreviations. Finally, we wanted to see to what extent these special characters affect the results of the OCR so we could see what aspects we need to improve.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

DICTIONARIES	TN CHARACTERS	TN ERRORS	% CORRECT	TN LETTERS WITH DIACRITICS	TN ERRORS IN DIACRITICS	% CORRECT	% ERRORS IN DIACRITICS VS. TN ERRORS
Cvetanović (2013)	1455	2	99,87	107	/	100	/
Đoković (2010)	2761	4	99,86	/	/	/	/
Boričić Tivranski (2002)	1232	2	99,84	45	/	100	/
Cvijetić (2014)	2791	17	99,39	30	/	100	/
Zlatković (2014)	3422	33	99,04	263	6	97,8	18,18
Stanić (1990–1991)	4394	62	98,59	263	16	94	25,80
Ristić (2010)	2938	43	98,54	312	25	92	58,13
Dalmacija (2017)	2047	30	98,53	193	15	92,2	50
Rajković Koželjac (2014)	3011	47	98,44	175	20	88,6	42,55
Dalmacija (2004)	2938	38	98,42	329	5	98,5	13,15
RSGV (2000–)	3566	79	97,74	161	14	91,3	17,72
Bašanović-Čečović (2010)	2853	61	97,86	355	35	90,1	57,37
Bukumirić (2012)	2563	72	97,19	256	6	97,7	8,33

Table 2: Accuracy of OCR processing

As it is shown in Table 2, three dictionaries have the highest accuracy percentage – 99,87% in Cvetanović (2013), Đoković (2010) and 99,86%, and 99,84% in Boričić Tivranski (2002). A mutual characteristic they all share is zero errors in characters with diacritics. In addition, one more dictionary is processed without errors in diacritics, Cvijetić 2014, making it a total of four.

When it comes to the total number of errors in diacritics, most of them are linked to characters in cursive. Dictionaries that have diacritics in cursive have the most mistakes in diacritics – Rajković Koželjac (2014) with 20 out of 175 total characters with diacritics (88,6% of accuracy), Bašanović Čečović (2010) with 35 out of 355 total (90,1%) and RSGV (2000)– with 14 out of 161 total (91,3%).

A specific type of error in characters with diacritics is present in most dictionaries – the letter о with any sort of diacritic is mistakenly read by the "SCyDia" application as the Cyrillic letter д. The most significant number of these errors is found in two dictionaries (Ristić 2010; Dalmacija 2017), where they form more than 50% of all errors in characters with diacritics.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

DICTIONARIES	TN CHARACTERS IN CURSIVE	TN ERRORS IN CURSIVE	TN ABBREVIA-TIONS	TN ERRORS IN ABBREVIATIONS
Cvetanović (2013)	/	/	75	/
Đoković (2010)	/	/	78	3
Boričić Tivranski (2002)	85	1	55	1
Cvijetić (2014)	798	17	228	5
Zlatković (2014)	755	12	298	6
Stanić (1990–1991)	1252	55	267	4
Ristić (2010)	828	1	75	/
Dalmacija (2017)	669	34	54	2
Rajković Koželjac (2014)	231	16	125	/
Dalmacija (2004)	820	1	83	/
RSGV (2000–)	148	1	452	20
Bašanović-Čečović (2010)	627	2	71	2
Bukumirić (2012)	483	14	152	17

Table 3: Accuracy of OCR processing additional data

Table 3 is providing further results obtained from processing the dictionaries in the "SCyDia" application.

What the results in the table are showing is that the presence (or lack) of cursive is crucial to the total percentage of errors, especially if cursive is combined with diacritics. Dictionaries with the highest percentage of errors (Bašanović Čečović 2010; Dalmacija 2017) have both characters in cursive and with diacritics. Similarly, dictionaries with the highest percentage of accuracy, such as Đoković (2010), Cvetanović (2013) don't have characters in cursive.

These results are similar to ones obtained by Polomac and Lutovac Kaznovac in their work with OCR for Serbian medieval manuscripts: "An extraordinarily high percentage of errors indicates that it is necessary to train a separate model for the automatic recognition of manuscripts written in cursive script" (Polomac/Lutovac Kaznovac 2021, p. 16). Although their system is trained to recognize manuscripts and Old Slavonic letters, it is interesting to see that cursive poses the biggest problem similarly to our results. It is also noteworthy to point out that the significan percentage of errors in their research are most frequently related to the blanks between words, superscript letters and titles, i. e. diacritics (ibid., pp. 23 f.).

4. Further plans

Once the transcribed text is manually corrected, we will place results in structured dictionary. We are currently developing an OntoLex schema that would be suitable for all the dictionaries and enable the smooth integration of various resources into one connected data structure. In the end, we want to create a web app with which some parts of the database would be accessible to the broader public, and some would require a license to access, depending on the

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

copyright of the dictionary. Also, the web app would allow a certain number of users to edit mistakes that may have remained after OCR and the scarce manual correction.

5. Conclusions

Today, when most dictionaries are being produced in digital form, it is essential not to lose sight of those that, for now, exist in paper form only and need to be transformed into a digital, computer-readable format. Breathing new life into non-digital lexicographic works requires a lengthy, multi-step process of retro-digitization. The end goal is to produce structured and indexed material that can be searched and integrated into various lexicographic projects, from scholarly dictionaries to more popular content. Still, in the case of the Serbian language, this end goal may look out of reach until some basic requirements are fulfilled. The presented "SCyDia" software solution is just one – but vital – step towards building up-to-date, multipurpose, and scientifically reliable digital linguistic resources for Serbian. "SCyDia" is developed as open-source software is available and it is available on GitHub at the following link: https://github.com/ilicv/Cyrilic_OCR.

References

Bašanović-Čečović, J. (2010): Rječnik govora Zete. Podgorica. [Vocabulary of Zeta (in Cyrillic)]

Boričić Tivranski, V. (2002): Rječnik vasojevićkog govora. Beograd. [Vocabulary of Vasojevići (in Cyrillic)]

Bukumirić, M. (2012): Rečnik govora severne Metohije. Beograd. [Dictionary of the north Metohia (in Cyrillic)]

Cristea, D./Pădurariu, C./Rebeja, P./Onofrei, M. (2020): From scan to text. Methodology, solutions and perspectives of deciphering old cyrillic Romanian documents into the Latin script. In: Knowledge, Language, Models, pp. 38–56. https://profs.info.uaic.ro/~dcristea/papers/Paper%20volume%20Bulgaria-Cristea_etAl.pdf (last access: 15-03-2022)

Cvetanović, V. (2013): Rečnik zaplanjskog govora. Gadžin Han. [Vocabulary of Zaplanje (in Cyrillic)]

Cvijetić, R. (2014): Rečnik užičkog govora. Užice. [Vocabulary of Užice (in Cyrillic)]

Dalmacija, S. (2004): Rječnik govora Potkozarja. Banja Luka. [Vocabulary of Potkozarje (in Cyrillic)]

Dalmacija, S. (2017): Rječnik govora Srba zapadne Bosne. Banja Luka. [Vocabulary of Serbian vernaculars of western Bosnia (in Cyrillic)]

Đoković, Lj. (2010): Rječnik nikšićkog kraja. Podgorica. [Vocabulary of the area of Nikšić (in Cyrillic)]

Hussain, S./Niazi, A./Anjum, U./Irfan, F. (2014): Adapting tesseract for complex scripts: An example for Urdu Nastalique. In 2014 11th IAPR International Workshop on Document Analysis Systems. IEEE, pp. 191–195.

Klyshinsky, E./Karpik, O./Bondarenko, A. (2020): A comparison of neural networks architectures for diacritics restoration. In: Recent Trends in Analysis of Images, Social Networks and Texts. AIST 2020. Communications in Computer and Information Science 1357, pp. 242–253.

Krstev, C./Stankovic, R./Vitas, D. (2018): Knowledge and rule-based diacritic restoration in Serbian. In: Computational Linguistics in Bulgaria. Proceedings of the Third International Conference (CLIB), Sofia, 28–29 May 2018. Sofia, pp. 41–51.

O'Brien, S./Haddej, D. B. (2012): Optical character recognition. Degree of Bachelor of Science. Worcester Polytechnic Institute.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Polomac, V./Lutovac Kaznovac, T. (2021): Automatic recognition of Serbian medieval manuscripts by applying the transcribus software platform: Current situation and future perspectives. In: Zbornik Matice srpske za filologiju i lingvistiku 64/2, pp. 7–26.

Prepis: http://www.prepis.org/ (last access: 01-03-2022).

Rajković Koželjac, Lj. (2014): Rečnik timočkog govora. Negotin. [Vocabulary of Timok (in Cyrillic)]

Raskovnik: http://raskovnik.org/ (last access: 01-03-2022).

RSGV (2000–): Rečnik srpskih govora Vojvodine. Novi Sad. [Vocabulary of Serbian vernaculars of Vojvodina (in Cyrillic)]

Rijhwani, S./Anastasopoulos, A./Neubig, G. (2020): OCR post correction for endangered language texts. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, pp. 5391–5942. https://aclanthology.org/2020.emnlp-main.478/ (last access: 15-03-2022)

Ristić, D. (2010): Rječnik govora okoline Mojkovca. Podgorica. [Vocabulary of the area of Mojkovac (in Cyrillic)]

Stanić, M. (1990–1991): Uskočki rečnik. Beograd. [Vocabulary of Uskoci (in Cyrillic)]

Stanković, R./Stijović, R./Vitas, D./Krstev, C./Sabo, O. (2018): The dictionary of the Serbian Academy: From the text to the lexical database. In: Lexicography in global contexts. Proceedings of the 18th EURALEX International congress, Ljubljana, 17–21 July. Ljubljana, pp. 941–949. https://euralex.org/publications/the-dictionary-of-the-serbian-academy-from-the-text-to-the-lexical-database/ (last access: 05-03-2022).

Stijović, R./Stanković, R. (2018): Digitalno izdanje Rečnika SANU: formalni opis mikrostrukture Rečnika SANU. In: Naučni sastanak slavista u Vukove dane 47/1, pp. 427–440.

Vitas, D./Krstev, C. (2015): Nacrt za informatizovani rečnik srpskog jezika. In: Naučni sastanak slavista u Vukove dane – Srpski jezik i njegovi resursi: teorija, opis i primene 44/3, pp. 105–116. [Blueprint for the computerized dictionary of the Serbian language (in Cyrillic)]

Zlatković, D. (2014): Rečnik pirotskog govora. Beograd. [Vocabulary of Pirot (in Cyrillic)]

Contact information

Velibor Ilić
The Institute for Artificial Intelligence Research and Development of Serbia, 21000 Novi Sad, Serbia
velibor.ilic@ivi.ac.rs

Lenka Bajčetić
Institute for the Serbian Language of SASA
lenka.bajcetic@gmail.com

Snežana Petrović
Institute for the Serbian Language of SASA
snezzanaa@gmail.com

Ana Španović
Institute for the Serbian Language of SASA
tesicana@gmail.com

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Acknowledgements

This paper was funded by the Ministry of Education, Science and Technological Development of the Republic of Serbia according to the Agreement No. 451-03-68/2022-14 concluded with the Institute for the Serbian Language of SASA.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Dorielle Lonke/Ilan Kernerman/Vova Dzhuranyuk

LEXICAL DATA API

Abstract This API provides data from various dictionary resources of K Dictionaries across 50 languages. It is used by language service providers, app developers, and researchers, and returns data as JSON documents. A basic search result consists of an object containing partial lexical information on entries that match the search criteria, but further in-depth information is also available. Basic search parameters include the source resource, source language, and text (lemma), and the entries are returned as objects within the *results* array. It is possible to look for words with specific syntactic criteria, specifying the part of speech, grammatical number, gender and subcategorization, monosemous or polysemous entries. When searching by parameters, each entry result contains a unique entry ID, and each sense has its own unique sense ID. Using these IDs, it is possible to obtain more data – such as syntactic and semantic information, multiword expressions, examples of usage, translations, etc. – of a single entry or sense. The software demonstration includes a brief overview of the API with practical examples of its operation.

Keywords API; lexical data; search; dictionary

1. Introduction

We present a Web RESTful API [1] that offers expansive lexical data originating from diverse lexicographic resources of K Dictionaries (KD) across 50 languages, including monolingual cores as well as bilingual pairs and numerous multilingual combinations. The target users include language service providers for their translation, localization and lemmatization tasks, application developers for language learning, games and word puzzles, and researchers in the academia and industry for enhancing NLP features and training machine learning models.

The idea for this API was initially conceived in the context of LDL4HELTA – Linked Data Lexicography for High-End Language Technology Application – a Eureka bilateral project involving KD and Semantic Web Company (2015–2017) that was aimed at offering dictionary content as linked data for NLP purposes ([2], [3]), and it continued to evolve and materialize as part of a Horizon 2020 project Lynx – Legal Knowledge Graph for Multilingual Compliance Services (2017–2021, [4]) for the development of a domain-specific multilingual knowledge management platform. In its alpha phase, the API was available online only for pre-approved users who received credentials for testing and provided input on the interface and functionality. Next (2018), it was launched in a public website, allowing access to registered users on a freemium basis. This launch served as a beta phase, in which conclusions were drawn regarding the usage pattern and engagement of registered users. Despite only minimal promotion, the API attracted interested users based on keyword search. In 2019, a market-ready version was released, including new user plans and an integrated payment processing system.

Prior to release, a comprehensive market research was performed to assess the availability of similar products and services available for commercial use. It was discovered that although several other renowned dictionaries offered access to their data in similar configurations, there were limited options for comprehensive lexical data APIs that provide full coverage of a lexicographic entry [5]. Many other services existed, offering well-structured results of linguistic data mining or machine-translated keywords [6]. These types of services

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

gained popularity in recent years, as computational methods for linguistics have become more prevalent. They offer an advantage of quickly collected data in great amounts that can be used for technical applications. However, this type of data is often automatically generated and as such it has limited use for more comprehensive linguistic research. Today, there are still only a handful of services providing reliable, human-curated data that can be used for linguistics as well as computational purposes, even more so for under-resourced languages.

The API returns data as JSON documents. A basic search result consists of an object containing partial lexical information on entries that match the search criteria, but further in-depth information is also available. It is possible to look for words with specific syntactic criteria, specifying the part of speech, grammatical number, gender and subcategorization, semantic and syntactic components, multiword expressions, and monosemous or polysemous entries.

In section 2 we describe the data resources and formats, section 3 presents the technical infrastructure and functionality, and the modelling is explained in section 4. section 5 presents our dissemination modes and conclusions.

2. Data

2.1 Resources

The API offers data from three different lexicographic resources, which are regularly updated and enriched, including:

(A) A series of extensive dictionary datasets for 25 languages, all adhering to the same macrostructure and microstructure and created from scratch (except two languages that rely on exterior resources, and which were adapted accordingly). Most of these sets are multi-layered, that is containing a monolingual base with translations in bilingual and multilingual levels. [7]

(B) A series of English bilingual learner's dictionaries including versions for 45 languages. [8]

(C) A legacy English monolingual dictionary. [9]

In addition, the API incorporates privately developed lists of inflected forms with morphology techniques that enable looking up word inflections and obtaining results from their main lemmas.

2.2 Format

Each of the three resources has its own data structure.

The first (A) consists of 25 lexical datasets, each representing a different source language and altogether including nearly a hundred bilingual versions. The entry in each one has two parts: headword block and sense block. The former provides details of alternative spelling and scripts, pronunciation, grammatical details, and inflected forms; the latter conveys the senses of the entry, each usually including a definition and related semantic labels (e.g. synonym, antonym, domain, register, etc.), example(s) of usage, (and multiword expressions,) and translation equivalents for each sense, example and expression. Some translations include additional information on grammatical gender and number, geographical usage, and

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

irregular inflected forms. The microstructure of expressions resembles that of the entry's senses and may also include various additional lexical details as described above, as well as be divided into different senses.

The second (B) stems from a single dataset that holds together the full English entries along with a translation for each sense in each of the other 45 target languages. The entry consists of a headword container and sense containers, which roughly include similar components as those of (A), with the exception of translations for the examples of usage.

The third (C) provides rich lexical details also including headword variants, geographical and biographical names with corresponding information, notes on language, spelling and grammar, and etymology.

3. Structure and functions

3.1 Infrastructure

The API backend relies on Elasticsearch [10], which enables efficient text searches and uses stemmers for several languages, making it possible to find words when searching their inflected forms. Stemmers create a stem word from the given word, although such stem does not have to be a valid word per se. For example, *argued* or *arguing* might be stemmed to the form *argu*. Stemming helps to find the appropriate headword and is a very convenient tool for searching.

The documents are stored as objects, whose structure is defined according to the mappings provided to Elasticsearch. Each data resource has its own mappings, and those in turn can differ based on language. Other object fields such as part of speech, grammatical gender, grammatical number, inflections (included as an integral part of the source data, usually as irregular forms, not those generated by language stemmers used by Elasticsearch), are used for the search.

The backend part is hosted on Amazon Web Services (AWS), while Elasticsearch is hosted on ElasticCloud. Both platforms are highly reliable and scalable, and have undergone massive testing and validation by our team.

3.2 Functionality

The API endpoint is located at [11]. There are two main methods for querying the API. The first, GET /search, allows the user to search for entries by specifying parameters such as language and headword text. This call returns a JSON object that contains an array of results, which is a list of entries that match the search criteria. Each entry in the result array contains the unique entry ID, the headword text and part of speech, and a list of its different senses, which, in turn, include their own unique identifiers and the definition text for disambiguation purposes. The lexical information provided by this call is partial.

A basic query requires headword text and a language code, but it is possible to further specify the search by adding optional parameters such as part of speech, grammatical number, gender or subcategorization, or narrowing the search to only monosemous or polysemous entries. The language codes mostly adhere to the ISO 639-1 code convention, but it is possible to obtain a list of all language codes by querying GET /languages. Since there can be multiple entries with the same headword text (differing in part of speech, for example), this

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

preliminary result allows the user to select the relevant option prior to obtaining the entire entry information.

The second method is GET /entries (GET /senses), which enables users to query the entire entry (sense) database across all languages, using the unique entry IDs obtained through the first method. This querying method posits that the user had already obtained the entry ID and returns various syntactic and semantic information, expressions, usage examples, translations, and more – of that single entry (or sense). The result object contains the following information:

- **id** (string) – the unique dictionary entry ID
- **source** (string) – the lexicographic resource from which the entry is taken
- **language** (string) – a two-character string that is the language code
- **version** (number) – the version of the dictionary the entry is taken from
- **related entries** (array of strings) – an array containing the IDs of the related entries
- **headword** (object/array of objects) – contains extensive phonetic and syntactic information of the headword(s)
- **senses** (array of objects) – contains an elaborate disambiguation of the headword into sense(s), including phonetic, syntactic and semantic information

The structure of the complete dictionary entry consists of two layers: the headword layer, which contains grammatical information pertaining to the entry regardless of its meaning, and the senses layer, which contains a semantic disambiguation of the word into its different meanings. Information about the usage of the word in context, such as example phrases, semantic category, register, synonyms and antonyms, and, of course, translations – which almost always depend on the particular meaning – is stored in the sense layer. To that extent, there exists a collection of senses, apart from the entry collection, within the API core, and they can be queried individually using the unique ID obtained in the initial search.

The API also includes two more complex functionalities designed to increase the number of results per search, *morph* and *analyzed.* These are Boolean parameters added to a GET / search query, and they expand the pool of results for a given word by including inflected forms or stems, disregarding diacritics and vocalization (for example in Arabic and Hebrew), and removing case-sensitivity (uppercase/lowercase). The parameters operate as follows:

- ***morph*** (boolean) – searches for the text in both headwords and inflections, including in our supplemental morphological lists. This is based on existing human-curated data and semi-automatically generated morphological lists, e.g., querying for the word *doors* will return the entry "door" (noun).
- ***analyzed*** (boolean) – this relies on a stemmer algorithm that strips words to their stem, disregarding diacritics and case (uppercase/lowercase), e.g., querying for the word *working* will return the entries *working* (adjective), *work* (verb), *work* (noun), *hard-working* (adjective), *working class* (noun), *work on* (verb), and any other entry with the stem *work* in its headword.

These functionalities were added after noticing that the previous search mechanism, which matched the exact text that was queried by the user to only the headwords in our data, was too restrictive. Opening the search for inflections and related entries consisting of the same lemma enabled a flexible query and proved economic for the users in terms of numbers of queries needed to obtain the desired results.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

4. Modelling

The structure that was chosen for the API response is convenient for integrating the API in external applications, while bearing similarity to the hierarchic structure of the original XML entry due to its nested form. The motivation behind this response structure was to encapsulate all the different components of a single dictionary entry. This allows the user to choose from the various information provided for a single headword, including the equivalents in the target languages. The JSON design was initially meant to facilitate the RDF conversion from XML to JSON-LD, which had several iterations from 2014 until its culmination in 2019 (cf. [12], [13], [14], [15], [16]), but eventually proved to be useful for the API users as well.

Upon examination of user behavior, it became evident that there were benefits for providing complete dictionary information. For one, it sets this API apart from other services that focus on one aspect of a headword, like translations, examples or grammatical information. This is much better utilized for linguistic research in which syntactic and semantic information are as relevant as translations or phrases, but also carry full potential for users who need specific information, which can be easily selected and filtered from the response in the user's application. Furthermore, the single-entry response structure enabled flexible search, as it provided the user with extensive information about a particular entry, and for that entry only. In this way, users are exposed to the entirety of information that is available per entry, without having to purchase a full corpus of words, and can select only those relevant for their needs and parse for the relevant information while still receiving the fullest extent of a dictionary entry.

The decision to provide a single entry at a time, but enable access to all of the entry components, stemmed from the thought that there is no added value in otherwise providing mass amounts of headwords or translations. Services offering large corpora of headwords or translations already exist, and a similar result can be achieved by more tech-savvy users by mining existing corpora and applying machine translation tools on their own. However, while the obvious advantage of large sets of headwords or translations is the amount of data a user receives, they usually lack the crucial information needed for linguistic research or language learning applications, and also might be reduced in quality. Our lexicographic resources are compiled based on frequency lists and automatic spellcheckers, but they undergo meticulous editing by human lexicographers who select the most relevant headwords, perform additional spelling revisions and grammatical corrections, and consolidate the headword list to a uniform configuration. More details, such as part of speech, grammatical gender and number, and geographical usage are then added, further enriching the dataset, and any additional information is allocated to the corresponding components in a systematic way. This ensures that the information provided inside the *Grammatical Gender* field, for example, is indeed grammatical gender information, and the correct one for that matter, as it has been added by a competent editor. The API structure allows the user to receive an abundance of information in one place, rather than later relying on separate resources to obtain further information for a headword, and to have the assurance that the available information has been manually curated. Instead of receiving huge lists of single values, users choose which words to access and receive the entirety of grammatical, syntactic, and semantic information related to that word. In that sense, we value quality over quantity.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

5. Dissemination and conclusions

The API was advertised mainly in inner circles of the lexicographic community, a couple of articles appeared in print and online [17], [18], and several hands-on workshops and software demonstrations were held in conjunction with lexicography conferences such as at EuraLex (2018, [19]) and eLex and AsiaLex (2019, [20], [21]), or the LTI summit (2019, [22]). Additionally, some SEO (search engine optimization) work was put into promoting the API and further market research pertaining to related keywords searched by potential users. During that process we discovered that public interest tended to focus more on dictionary APIs, particularly bilingual or multilingual ones, rather than on lexical data. To that extent, we invested in highlighting the lexicographic structure of the data alongside its modularity and functionality as broader lexical data to be used in other linguistics contexts. In terms of applications, it was evident that this type of product was sought after by various users, and the complete dictionary entry format proved to be useful in multiple contexts, further reinforcing the decision to model API responses as complete dictionary entries. Our user base consisted mainly of businesses who relied on the data for their own service, but there were occasionally individual users who utilized the data for their own research. This is congruent with our B2B, *Data as a Product*, business model data, but also showed that this type of dictionary data can be useful for smaller-scale products, as its modular character of a single-entry response allows for scaling up or down at the user's end. The provision of single entries per search was another strength point for the API, as it eliminated the need to acquire entire corpora or datasets and let the user select only relevant information (see section 4). For dissemination purposes, this was an advantage, as it allowed us to highlight the benefits of a full dictionary entry as a product with intrinsic value. This also allowed us to reach multiple market sectors: those interested in specific components, such as word lists or translations, could rely on these services to extract the information relevant to them; those seeking an array of lexicographic information were able to obtain a wealth of components in one place, rather than from separate services.

Future work that was considered is further developing the search mechanism to allow searching in particular components of the dictionary, such as examples, multiword expressions, or translations to a selected target language. This entails adding collections of particular components across dictionaries, which shifts the hierarchy from the traditional lexicographic sorting to a component-based filter applied on all datasets as one. A point in favor of this is the discovery that many users were only interested in certain aspects of a dictionary entry, as reflected by their queries. This would be another step in the direction of a lexical data API, as the extensive lexical information currently being offered would be accessible through a modular search structure that is not limited by the current structure of a dictionary entry result. As of now, the entirety of the lexical data is provided per dictionary and per entry, and particular information can be easily accessed by the user on their end by parsing the entry as they please. Adding other search mechanisms would not be changing the data but rather the means of delivery, possibly facilitating the usage for some users who are looking for non-lexicographic language solutions. This type of structural change reflects a shift from providing dictionary data to offering pieces of lexical information as per the user's request, without imposing the lexicographic configuration. While we are interested in exploring this direction, it is evident that there is still high interest in a dictionary API, with all the benefits of the complete lexicographic structure provided as a whole.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

References

[1] https://api.lexicala.com/.

[2] Kaltenböck, M./Kernerman, I. (2017): Introducing LDL4HELTA: Linked data lexicography for high-end language technology application. Kernerman Dictionary News 25. https://lexicala.com/review/#no25.

[3] Turdean, T./Joshi, S. (2017): Triplifying a dictionary: some learnings. In: Kernerman Dictionary News 25. https://lexicala.com/review/#no25.

[4] https://lynx-project.eu/.

[5] https://programmableweb.com/category/dictionary/api.

[6] https://programmableweb.com/news/10-popular-apis-words/brief/2021/03/12.

[7] GLOBAL series: Arabic, Chinese (Simplified and Traditional), Czech, Danish, Dutch, English, French, German, Greek, Hebrew, Hindi, Italian, Japanese, Korean, Latin, Norwegian, Polish, Portuguese (Brazilian and European), Russian, Spanish, Swedish, Thai, Turkish.

[8 PASSWORD English semi-bilingual dictionaries.

[9] Random House Webster's college dictionary.

[10] https://en.wikipedia.org/wiki/Elasticsearch.

[11] https://lexicala1.p.rapidapi.com.

[12] Klimek, B./Brümmer, M. (2015): Enhancing lexicography with semantic language databases. Kernerman Dictionary News 23. https://lexicala.com/review/#no23.

[13] Bosque-Gil, J./Gracia, J./Gómez-Pérez, A. (2016): Linked data in lexicography. In: Kernerman Dictionary News 24. https://lexicala.com/review/#no24.

[14] Aguado-de-Cea, G./Montiel-Ponsoda, E./Kernerman, I./Ordan, N. (2016): From dictionaries to cross-lingual lexical resources. In: Kernerman Dictionary News 24. https://lexicala.com/review/#no24.

[15] Bosque-Gil, J./Gracia, J./Montiel-Ponsoda, E. (2017): Towards a module for lexicography in OntoLex. In: Kernerman Dictionary News 25. https://lexicala.com/review/#no25.

[16] Lonke, D./Bosque-Gil, J. (2020): Applying the OntoLex-lemon lexicography module to K Dictionaries' multilingual data. In: Kernerman Dictionary News 28. https://lexicala.com/review/2020/lonke-bosque-gil-ontolex-lemon-lexicog/.

[17] Alper, M. (2017): KD API. In: Kernerman Dictionary News 25. https://lexicala.com/review/#no25.

[18] Kernerman, I./Lonke, D. (2019): Lexicala API: a new era in dictionary data. In: Kernerman Dictionary News 27. https://lexicala.com/review/#no27.

[19] https://euralex2018.cjvt.si/.

[20] https://elex.link/elex2019/.

[21] https://asialex2019.istanbul.edu.tr/en/_.

[22] https://www.lt-innovate.org/award.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Contact information

Dorielle Lonke
K Dictionaries
dorielle@kdictionaries.com

Ilan Kernerman
K Dictionaries
ilan@kdictionaries.com

Vova Dzhuranyuk
K Dictionaries
vova@kdictionaries.com

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Takahiro Makino/Rei Miyata/Seo Sungwon/Satoshi Sato

DESIGNING AND BUILDING A JAPANESE CONTROLLED LANGUAGE FOR THE AUTOMOTIVE DOMAIN

Toward the development of a writing assistant tool

Abstract In this paper, we propose a controlled language for authoring technical documents and report the status of its development, while maintaining a specific focus on the Japanese automotive domain. To reduce writing variations, our controlled language not only defines approved and unapproved lexical elements but also prescribes their preferred location in a sentence. It consists of components of a) case frames, b) case elements, c) adverbial modifiers, d) sentence-ending functions, and e) connectives, which have been developed based on the thorough analyses of a large-scale text corpus of automobile repair manuals. We also present our prototype of a writing assistant tool that implements word substitution and reordering functions, incorporating the constructed controlled language.

Keywords Japanese controlled language; corpus-based lexicon building; variation management; writing support tool; automotive domain

1. Introduction

The production process of technical documents for industrial products, such as automobile repair manuals, usually involves many writers and editors. This induces writing variations, which might not only degrade the searchability of document content for readers, but also the reusability of past text for writers. These variations may also have a negative impact on translation memory tools, degrading the reusability of past translations.

To reduce such variations, it is crucial to use a properly designed controlled language in combination with a writing assistant tool. Controlled languages restrict the syntax and/or lexicon of a certain natural language (Kittredge 2003; Kuhn 2014). The syntactic restrictions are usually defined through writing rules, such as 'write short and clear sentences' (ASD 2021). Although such rules provide a general guide to writing, some may not concretely indicate how to compose a sentence. The lexical restrictions take the form of a controlled lexicon, which is a list of approved words. The lexicon becomes more useful if unapproved words are linked to approved words (Warburton 2014). Another challenge is the provision of detailed descriptions of word usage. For languages with flexible word order, such as Japanese, the regulation of word locations in sentences is important when improving textual consistency.

While many English controlled languages have been proposed and used in practice (Kuhn 2014), the development of Japanese controlled languages has not advanced sufficiently. Current approaches to Japanese controlled languages for writing purposes mainly address syntactic restrictions (e. g., Japan Technical Communicators Association 2016; Japio 2018). Furthermore, the tools for assisting Japanese controlled writing are scarce (Miyata et al. 2016).

Against this backdrop, domain-specific lexical restrictions are needed to properly manage the writing variations. Miyata/Sugino (2020) reported the building process of a Japanese controlled lexicon for the automotive domain, specifically focusing on the verbs and their

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

case orders. To further cope with various writing variations, we need to extend the scope to cover other lexical elements necessary for writing sentences. In this study, therefore, we design and build a Japanese controlled language for writing technical documents in the automotive domain that covers a wide range of linguistic elements. We also introduce our prototype tool designed to help writers reduce various types of writing variations.

In section 2, we propose our controlled language with its design principle, components, and the general methodology to build it. We then present the three of the components in sections 3–5, respectively, showing the detailed building process and results. In section 6, we introduce a prototype of our writing assistant tool that implements part of the controlled language. Finally, we conclude this paper with implications for future work in section 7.

2. Design of controlled language

2.1 Principle

To avoid writing variations, our controlled language should comply with the principle 'one meaning/function should correspond to one form'. Some controlled languages, including ASD-STE 100 (ASD 2021), are designed to comply with the reverse ('one form should correspond to one meaning/function') as well because polysemy might lead to the ambiguity of reading and hinder the readers' understanding. Although we did not count this as a requirement, we aimed to achieve this when building a controlled language.

As mentioned in section 1, for the purpose of controlled writing, it is effective to include unapproved words in the lexicon and link them to approved words (Warburton 2014, 2021). ASD-STE 100, for example, regulates the use of an unapproved verb 'delete' and provides its alternatives, namely, 'disconnect', 'disengage', and 'remove', each of which is defined to have a unique approved meaning. The comprehensive prescriptions of such linkages between unapproved and approved words are possible if the target domain is sufficiently specified. Like ASD-STE 100, which originally focused on aerospace maintenance documentation, our controlled language is intended for a specific domain. Thus, we reasonably assume that we can widely define the unapproved words in addition to approved ones.

Another important principle of our controlled language is that syntactic restrictions are incorporated in the lexical components. Previous controlled lexica often do not explicitly specify the detailed syntactic information. ASD-STE 100, for example, provides an approved word 'remove' with its part-of-speech information ('v'), approved meaning ('To take or move something away from its initial position'), and approved example ('REMOVE THE INDICATOR FROM THE PANEL.') (ASD 2021, p. 2-1-R9). However, for languages with flexible word order, more detailed information might be needed to avoid writing variations. In Japanese, for example, the following two sentences are grammatically correct:

(1) インジケータを パネルから 取りはずす。/ *Injiketa o paneru kara torihazusu.*

(2) パネルから インジケータを 取りはずす。/ *Paneru kara injiketa o torihazusu.*
(Translation: Remove the indicator from the panel.)

In this example, the case order for the verb *torihazusu* (remove) is swapped between (1) and (2), without changing the sentence's meaning. To control such variations, it would be effective to provide preferred word order information in the entries of the lexicon.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

2.2 Components

Although Miyata/Sugino (2020) already compiled a controlled lexicon of verbs, they did not cover other parts of speech, such as nouns and adverbs. Therefore, to control the extensive range of writing variations, we designed a controlled language that consists of the following five components:

(a) **case frames:** verbs with canonical orders for their argument slots (Miyata/Sugino 2020)
(b) **case elements:** nouns or noun phrases that can fill argument slots
(c) **adverbial modifiers:** adverbs or adverbial phrases with their preferred locations
(d) **sentence-ending functions:** sequences of functional words attached to sentence-ending main verbs
(e) **connectives:** conjunctions and conjunctive phrases that indicate inter-clause or inter-sentence relationships

Figure 1 shows an example of a Japanese sentence annotated using our controlled language. Most sentences in our target documents can be broken down into elements derived using the five components.

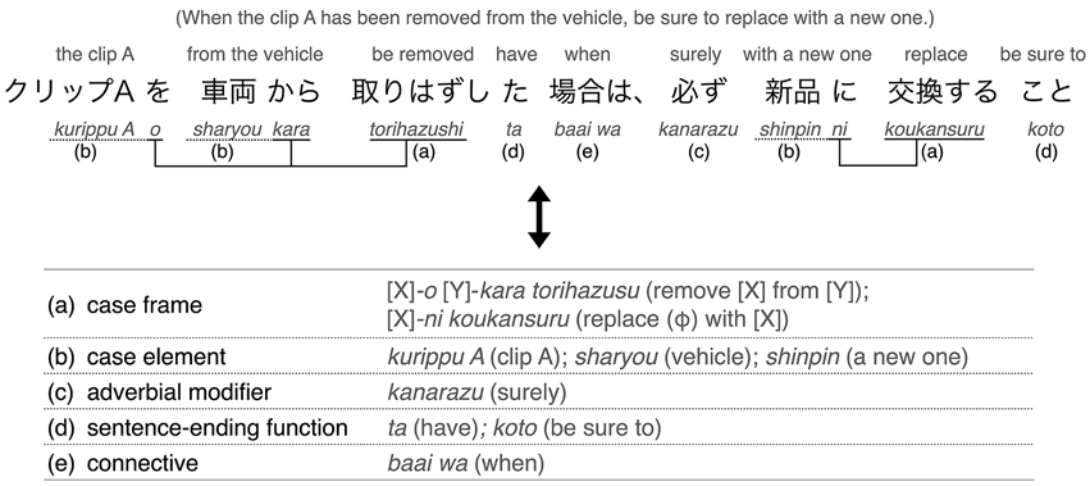


Fig. 1: Sentence annotated with five types of elements defined in our controlled language

It is significant that each entry in these components is not only a single word but also a sequence of words. For example, the connective *baai wa* (when) in Figire 1 is composed of a noun *baai* and a particle *wa*, and the combination of the two words can be regarded as a basic operational unit when writing sentences. While the existing controlled lexica tend to register single words as entries, we flexibly define the linguistic spans of lexicon entries with the purpose of controlled writing assistance.

Here, we specifically provide the details of components (c)–(e) in sections 3–5, respectively, which have not been sufficiently investigated in previous studies.[1]

[1] The component (a) was compiled by Miyata/Sugino (2020). The component (b) virtually means the terminology of the target domain. The terminology management for controlled authoring has been discussed by Warburton (2021).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

2.3 Methodology for controlled language building

We adopted a corpus-based method to build a controlled language. The corpus was first constructed by extracting sentences from 17 sets of automobile repair manuals provided by Toyota Motor Corporation. A total of 1,053,111 sentence tokens (158,383 sentence types) were collected (henceforth, **Corpus-Org**). As the Corpus-Org includes complex/compound sentences with more than one clause, we then decomposed them into simpler sentences using a Japanese sentence splitting tool (Kato et al. 2020a) and obtained 1,428,381 sentence tokens (159,816 sentence types) (henceforth, **Corpus-Simple**).

We built each component using the following steps: i) comprehensively collecting instances from the corpus; ii) grouping the instances in terms of meaning/function; iii) defining approved and unapproved types for each group; and iv) specifying the preferred usage of the approved types, if any.

In step ii), we specifically examined the *interchangeability* of instances in actual sentences in the corpus. The following examples shows the interchangeability of adverbial modifiers.

(3) 空気が<u>極力</u>入らないように張り付ける / *Kuuki ga <u>kyokuryoku</u> hairanai youni haritsukeru.*

(3') 空気が<u>なるべく</u>入らないように張り付ける / *Kuuki ga <u>narubeku</u> hairanai youni haritsukeru.*

(Translation: Attach so that air does not enter <u>as much as possible</u>.)

Here we observe that the word *kyokuryoku* (as much as possible) can be replaced with the word *narubeku* without changing the meaning of sentence (3). It should be noted that synonymous words obtained from general thesauri do not guarantee their interchangeability in the target domain text. In addition, words that are judged as interchangeable may not be regarded as synonymous in a strict sense.

In steps iii) and iv), to define the approved types or preferred usage, we chiefly referred to frequency information; the more frequently observed in the corpus, the more likely to be specified as approved/preferred. Since the previously-authored documents are primarily referred to by writers as 'models' in the writing process, it is reasonable to use the frequency information in the target corpus as significant evidence for defining approved words and preferred usage.

3. Lexicon of adverbial modifiers

The adverbial modifiers were divided into two types: fixed and variable. The former includes adverbs, such as 必ず/*kanarazu* (surely), and adverbial phrases, such as 同量ずつ/*douryouzutsu* (by an equal amount). The latter includes adverbial phrases with a slot, such as 約[X]分間/*yaku* [X] *funkan* (for approximately [X] minutes).

3.1 Building procedure

To widely collect instances of adverbial modifiers, we first manually investigated the 200 most frequent sentences in the Corpus-Org and defined heuristic rules to extract instances based on the parsed results of Japanese sentence analysis tools, Juman++ (Tolmachev/Kawahara/Kurohashi 2018) and KNP (Kawahara/Kurohashi 2006). Table 1 shows the formulated linguistic rules; they are defined based on the types of last morpheme (if it is a particle, the

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

second last) of each modifier of the main verb. Using these rules, we automatically collected 3,922 types of adverbial modifier candidates.

No	Last morpheme type	Extracted example
1	adverb	ゆっくり/*yukkuri* (slowly)
2	continuative form of adjective	無理に/*murini* (by force)
3	temporal noun that can be used as an adverb	再度/*saido* (again)
4	suffix to make an adverb: 的に/*tekini*, めに/*meni*	定期的に/*teiki-tekini* (periodically)
5	temporal suffix: 後/*go* (after), 中/*chu* (while), 前/*mae* (before), 時/*ji* (when)	作業前に/*sagyou mae-ni* (before proceeding with work)
6	numeral suffix: 回/*kai* (time), 秒間/*byoukan* (second) ,秒/*byou* (second), 分間/*funkan* (minute), 分/*fun* (minute), %, 種類/*syurui* (kind), 以上/*ijou* (more than), ずつ/*zutsu* (one by one), 程度/*teido* (about), 程/*hodo* (about)	15秒間/*15 byoukan* (for 15 seconds)

Table 1: Linguistic rules to extract adverbial modifiers

We then manually classified them into fixed and variable types. The instances extracted by rules 5 and 6 in Table 1 were mostly variable types. We identified a variable span in each instance and abstracted it to form a variable type, such as [X] *maeni* (before [X]) and [X] *byoukan* (for [X] minutes), where [X] indicates a variable. Each variable can be substituted by a noun (phrase), an adjective, or a numerical value.

The approved and unapproved elements were defined based on the procedures described in section 2.3. As adverbial modifiers can be flexibly inserted into sentences, we also defined a preferred insertion location for each approved element mostly based on the frequency information. The location options are as follows: 1) at the front of the sentence; 2) at the front of the clause; 3) immediately before the verb; 4) after the topical marker は/*wa*; 5) before the nominative case が/*ga*; and 6) after the nominative case が/*ga*.

3.2 Results

The statistics of the constructed lexicon of adverbial modifiers are presented in Table 2 for fixed types and in Table 3 for variable types. In this paper, we use a right arrow '→' to indicate the link from an unapproved element to an approved one.

For the fixed types, we defined 235 approved types and 38 unapproved types, which cover 63,402 and 1,713 tokens in the Corpus-Simple, respectively. This means that about 2.6% (1,713/65,115) of fixed adverbial modifiers in the corpus can be regarded as variations.

For the variable types, we defined 75 approved types and 64 unapproved types, which cover 73,976 and 6,719 tokens in the Corpus-Simple, respectively. As we used the automobile repair manuals as a source for lexicon building, various types of adverbial modifiers regarding *quantity* and *time* were collected, which are indispensable for describing the repair operations.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Category		Example entry	Approved		Unapproved	
Level 1	Level 2		# Type	# Token	# Type	# Token
manner		*yukkurito* →*yukkuri* (slowly), *kinnit-suni*→*kintouni* (uniformly), *kakujitsuni* (securely)	110	21,779	12	660
degree		*ooyoso*→*hobo* (almost), *tashou*→*sukoshi* (slightly), *sarani* (kanji)→*sarani* (kana) (further)	17	2,431	3	92
aspect	proximity	*tadachini*→*suguni* (immediately), *soku*→*suguni* (immediately), *mada* (still)	2	381	3	241
	continua-tion	*sukoshizutsu*→*jojoni* (gradually), *yuruyakani*→*jojoni* (gradually), *ichijitekini* (temporary)	9	2,937	2	119
	repetition	*futatabi*→*saido* (again), *saido-hajimekara* (again from the beginning)	2	3,971	1	143
	order	*ittan* (kanji)→*ittan* (kana) (for a while), *sonoatoni*→*sonogo* (then), *mazuwa*→*mazu* (first)	21	5,145	5	104
	frequency	*jouji*→*tsuneni* (always), *taezu*→*tsuneni* (always), *teikitekini* (at regular intervals)	6	2,100	2	37
emphasis		*kanarazu* (surely), *kesshite* (never), *zettai*→*zettai-ni* (never)	3	18,475	1	132
others		*suubyou*→*suubyoukan* (for several seconds), *tochude* (halfway through), *sorezore* (respectively)	65	6,183	9	185
		Total	235	63,402	38	1,713

Table 2: Statistics of the lexicon of adverbial modifiers (fixed type)

Category	Example entry	Approved		Unapproved	
		# Type	# Token	# Type	# Token
manner	[X] *to douyou ni* (in the same way as [X]), [X] *to issho ni* (with [X]), [X] *chokuzen de* (just before [X])	3	605	0	0
quantity	[X] *byou ijou* → *sukunakutomo* [X] *byoukan* (for at least [X] seconds), [X] *fun* → [X] *funkan* (for [X] minutes), [X] t o [X] *kai* ([X] to [X] times)	37	16,953	54	6,601
time	[X] *chokugo wa* (immediately after [X]), [X] *ji nado de* → [X] *ji nado ni* (when [X]), [X] *chu de* →[X] *chu ni* (during [X])	31	54,077	10	118
order	[X] *kara jun ni* (from [X] in order), [X] *no* [Y] *jun ni* (in [Y] order of the [X])	4	2,341	0	0
	Total	75	73,976	64	6,719

Table 3: Statistics of the lexicon of adverbial modifiers (variable type)

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

4. Lexicon of sentence-ending functions

The sentence-ending functions are word sequences that can be attached to the main verbs for adding various functional information (Kato/Miyata/Sato 2020b). For example, the phrase 表示されない場合がある/*hyoujisa-re nai baai ga aru* (**may not be** display**ed**) has three categories of functions—passive voice (*re*), negation (*nai*), and possibility modality (*baai ga aru*). In terms of controlled writing, each function should be expressed in the same form; for example, the possibility modality should be *baai ga aru* instead of *koto ga aru*.

4.1 Building procedure

As mentioned in section 2.2, the unit of a function does not necessarily correspond to a single word; for example, the possibility modality *baai ga aru* is a combination of a noun, particle, and verb. To identify appropriate spans of functions, we used a Japanese sentence-ending analyser Panzer (Sano/Miyata/Sato 2020), which is based on a domain-specific language for Japanese sentence composition (Sato 2020). We first analysed all the sentences in the Corpus-Simple using Panzer and obtained a set of sequences of sentence-ending functions. We then decomposed the 150 most frequent sequences to minimal units of functions and identified 31 function types. These function types were categorised, and approved/unapproved functions were defined based on the method presented in section 2.3.

4.2 Results

Table 4 shows the statistics of the constructed lexicon of sentence-ending functions. The category level 2 can be regarded as the parameters for the category level 1. For example, the voice can be specified by selecting one of the options: active,[2] passive, and causative. Importantly, for most of the categories in level 2, only a single approved element is defined, such as *koto ga dekiru* for potential modality. One of the exceptions is the possibility modality, which includes three approved elements: *baai ga aru*, *kanousei ga aru*, and *osore ga aru*. The examples from the corpus are presented below with their English translations:

(4) 待ち時間が発生する場合がある / *machijikan ga hassei suru baai ga aru*
(Translation: waiting time may be required)

(5) ダイアグを出力する可能性がある / *Daiagu o shutsuryokusuru kanousei ga aru*
(Translation: a DTC may be output)

(6) 不具合が発生するおそれがある / *fuguai ga hasseisuru osore ga aru*
(Translation: this may cause a malfunction)

As the three elements are translated into the same functional word 'may' in English, it is possible to unify them into a single approved form (e.g., *baai ga aru*). Nevertheless, since each expression has a unique nuance, i.e., 'there is a case' (*baai ga aru*), 'there is a possibility' (*kanousei ga aru*), and 'there is a risk' (*osore ga aru*), we retain them as approved elements for the sake of the expressivity of the controlled language.

[2] Although the active voice is an unmarked element in Japanese, we included it in Table 4 to explicitly show the distribution of voice types. The same applies to the affirmative polarity.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Category		Elements	Approved		Unapproved	
Level1	Level2		# Type	# Token	# Type	# Token
voice	active	[unmarked]	1	1,384,844	0	0
	passive	*reru*/*rareru*	1	43,537	0	0
	causative	*seru*/*saseru*	1	49,651	0	0
polarity	affirmative	[unmarked]	1	1,360,063	0	0
	negative	*nai*	1	68,318	0	0
modality	obligation	*hitsuyou ga aru*, *nakereba naranai* ⟶ *koto*	2	26,502	1	155
	potential	*koto ga dekiru*	1	6,651	0	0
	possibility	*baai ga aru*, *kanousei ga aru*, *koto ga aru* ⟶ *baai ga aru*, *osore ga aru* (kanji) ⟶ *osore ga aru* (kana)	3	32,130	2	5,490
	tendency	*yasui*	1	831	0	0
	trial	*te-miru*	1	280	0	0
	interrogative	*ka* ⟶ [remove]	0	0	1	198
	determination	*koto ni naru*	1	133	0	0
	permission	*te-yoi*	1	108	0	0
	request	*te-kudasai* ⟶ *koto* (obligation)	0	0	1	74
aspect	state	*te-aru* ⟶ *te-iru*	1	151,202	1	616
	perfect	*te-shimau*	1	1,570	0	0
	preparation	*te-oku*	1	3,828	0	0
	continuation	*te-iku* ⟶ [remove]	0	0	1	214
change		*naru*, *you ni suru*	2	5,093	0	0
completion		*ta*	1	1,963	0	0
parallel	illustration	*tari-suru*	1	1,100	0	0
honorifics	polite	*masu* ⟶ [remove]	0	0	1	495
	humble	*te-itadaku*	1	233	0	0
emphasis		*mo*	1	374	0	0

Table 4: Statistics of the lexicon of sentence-ending functions

5. Lexicon of connectives

In this study, we defined a *connective* as an expression that is located in between clauses (or sentences) to indicate their relationship.[3] We can distinguish two types of connectives: those located at the beginning of the clause (henceforth, clause-beginning connectives); and those located at the end of the clause (henceforth, clause-ending connectives).

[3] Although connectives are also used to combine various elements other than clauses, such as nouns and verbs, in this paper, we focus on clause-level connections.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

5.1 Building procedure

Linguistically, the clause-beginning connectives generally correspond to conjunctions and conjunctive adverbs. We thus used the Japanese morphological analysis tool Juman++ to widely extract these parts of speech as candidates of clause-beginning connectives.[4]

To collect candidates of clause-ending connectives, we identified coordinate and subordinate clauses in the Corpus-Org using Juman++ and KNP and extracted the connectives. As we discovered that the coordinate or adverbial clauses directly identified by the tools are not sufficient for our purpose, we added the pattern of an attributive clause with an adverbial parent element whose attributive relation is 'external' (Teramura 1975–1978).

We manually excluded the irrelevant candidates and controlled variations to define approved and unapproved elements based on the procedures described in section 2.3.

5.2 Results

Table 5 shows the statistics of clause-beginning types of connectives, while Table 6 shows that of clause-ending types. These types are categorised partly based on the typology of Japanese conjunctions by Ishiguro (2016).

It is notable that 37% of clause-ending connective tokens were deemed unapproved variations. The following types of variations were identified and controlled.

- Synonyms: e. g., よって/*yotte* → したがって/*shitagatte* (therefore)
- Character variations: e. g., 時/*toki* (kanji) → とき/*toki* (kana) (when)
- Post-positional particle variations: e. g., 場合/*baai* → 場合は/*baai wa* (in case)

Category	Example	Approved		Unapproved	
		# Type	# Token	# Type	# Token
resultative	*shitagatte* (kana) → *shitagatte* (kanji) (therefore), *yotte* → *shitagatte* (kanji) (therefore), *konotame* (for this reason)	4	625	2	135
adversative	*gyakuni* (conversely), *shikashi* (but)	2	96	0	0
parallel	*mata* (kanji) → *mata* (kana) (also), *katsu* (kana) → *katsu* (kanji) (besides), *sarani* (in addition)	6	4,915	2	23
contrast	*matawa* (kanji)→*matawa* (kana) (or), *aruiwa*→*matawa* (or), *moshikuwa*→*matawa* (or)	1	124	3	40
paraphrase	*tsumari* (in other words)	1	49	0	0
example	*tatoeba* (kana) → *tatoeba* (kanji) (for example)	1	31	1	2
addition	*nao* (in addition), *tadashi* (kanji) → *tadashi* (kana) (however)	2	609	1	398
	Total	17	6,449	9	598

Table 5: Statistics of the lexicon of connectives (clause-beginning types)

[4] This process was conducted in parallel with the lexicon building process for adverbial modifiers described in section 3.1 because the target linguistic elements overlap with each other.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Category			Example	Approved		Unapproved	
Level1	Level2	Level3		# Type	# Token	# Type	# Token
coordinate	resultative	and	V-*te*→V (continuative form)	1	218,165	1	199,945
	illustration	such as	*tari, nado*	2	6,737	0	0
	contradictory	but	*ga*	1	4,912	0	0
	cumulation	in addition	*ue*	1	17	0	0
subordina-te	time	when	*toki* (kanji)/*toki wa/sai/sai wa/ jiten de*→*toki* (kana), *sai ni/ toki ni* (kanji)→*toki ni* (kana), *sai niwa, toki niwa*→*toki* (kana)/*toki ni* (kana)	2	15,882	11	21,488
		each time	*tabi ni/goto ni* (kanji)→*goto ni* (kana)	1	115	2	63
		with	*to tomo ni* (kana)→*to tomo ni* (kanji)	1	222	1	119
		before	*mae niwa*→*mae ni, mae wa*	2	17,668	1	80
		after	*nochi/nochi ni/ato* (kana)/*ato ni/ato de* (kanji)→*ato* (kanji), *ato wa* (kanji), *kara*	3	23,563	6	1,849
		then	*ue de* (kana)→*ue de* (kanji)	1	396	1	215
		while	*aida wa, uchi ni*	2	1,623	0	0
		until	*made*	1	9,192	0	0
	condition	in case	*baai ni*→*baai, baaini wa/toki/ toki wa/toki ni/toki niwa/sai wa/sai niwa*→*baai/baai wa*	2	76,611	2	5,397
		if	*ba/naraba/nonara/ nodeareba*→*baai/baai wa, tara*→*ato* (kanji)/*toki* (kana)/ *baai/baai wa*	1	32,203	5	2,449
		if it seems	*you nara/you deareba*→*baai/ baai wa*	0	0	2	403
		though	*mo*	1	4,435	0	0
		as long as	*kagiri* (kana)→*kagiri* (kanji)	1	269	1	9
	methond	by	V (continuative form)→V-*te/ koto de*	1	7,019	0	0
	attendant circumstan-ces	with -ing	*tsutsu*→*nagara/mama*	2	9,597	1	27
		without -ing	*zuni*	1	3,337	0	0
	state	in the state	V (continuative form)→V-*te/ joutai de*	1	7,364	0	0
	purpose	in order to	*you/you ni* (kanji)→*you ni* (kana), *tame* (kanji)/*tame ni*→*tame* (kana), *tame niwa*→*niwa*	2	10,083	6	3,375
	reason	because	*tame*→*node*	1	3,100	2	37,507
		(structure)	*kekka*→V (end-form)	0	0	1	196

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/ Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Category			Example	Approved		Unapproved	
Level1	Level2	Level3		# Type	# Token	# Type	# Token
	contradictory	but	*noni*	1	243	0	0
	extent	to the extent	*teido*→*teido ni*, *hodo* (kana)→*hodo* (kanji)	2	208	2	62
	restrict	just	*dakede*	1	28	0	0
			Total	37	452,989	43	273,427

Table 6: Statistics of the lexicon of connectives (clause-ending types)

6. Writing assistant tool

Implementing the controlled language components mentioned in sections 3–5, we are developing a writing support tool to help writers and editors compose controlled sentences. Figure 2 presents the prototype interface of our tool. Similar to existing controlled language checkers (e.g., Bernth/Gdaniec 2001; Mitamura et al. 2003; Miyata et al. 2016; Nyberg/Mitamura/Huijsen 2003), our tool detects unapproved/non-preferred elements in the input text, suggests candidates of approved/preferred elements, and corrects the target segment when the user selects one of the candidates.

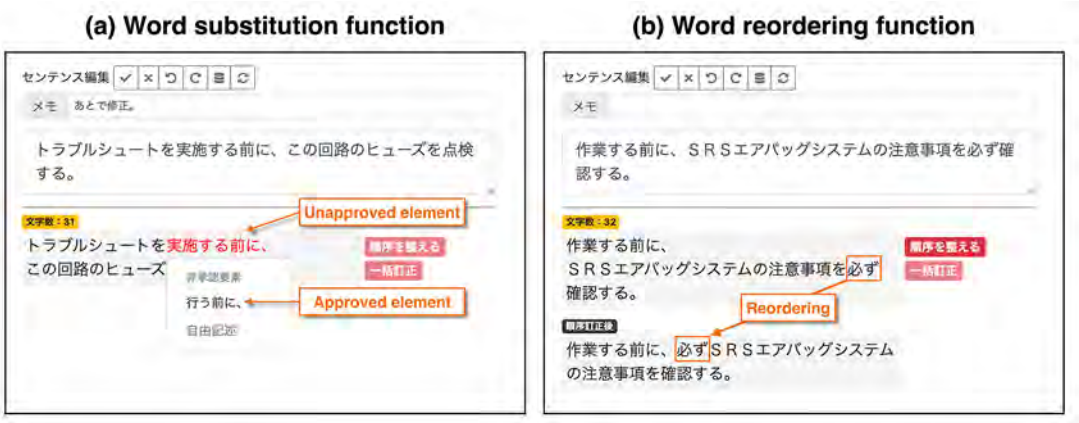


Fig. 2: Prototype interface of our writing assistant tool

The word substitution function in Figure 2 (a) is helpful to reduce the word type variations. This function can be implemented by using the pairs of unapproved and approved elements directly obtained from our controlled language components. However, if the input text includes elements that are not registered in the controlled language, we need to specify which approved elements should be suggested in an ad hoc manner. To search for an approved word that is contextually interchangeable to the unregistered element, it would be effective to use the similarity of word vectors obtained from contextual embedding models, such as ELMo (Peters et al. 2018) and BERT (Devlin et al. 2019).

The word reordering function in Figure 2 (b) can detect the word order variations and suggest the preferred order, based on the location information prescribed in our controlled language. The example in the figure illustrates the suggestion of the appropriate location of

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

an adverbial modifier 必ず/*kanarazu* (surely), the preferred location of which is defined as 'at the front of the clause'. This function is novel as existing controlled language checkers rarely support the correction of word order variations.

7. Conclusion and outlook

In this paper, we reported our attempt to design and build a Japanese controlled language that is intended to support controlled writing of automotive technical documents. The controlled language defines approved and unapproved lexical elements, extensively covering a) case frames, b) case elements, c) adverbial modifiers, d) sentence-ending functions, and e) connectives. The key principles are that it contains linkages between unapproved and approved elements and that it defines preferred word orders for approved elements. Our controlled language components have been constructed through the comprehensive analysis of a large-scale text corpus of automobile repair manuals. While we assume that our controlled language can widely cover the automotive domain, we plan to verify its applicability to other document types in the domain and continuously expand its coverage.

We assume that our methodology to build a controlled language is generally applicable to other domains outside the automotive domain if a sufficiently large corpus is available. While the sizes of components regarding functional words (i. e., d) sentence-ending functions and e) connectives) can be limited, those of the other components (i. e., a) case frames, b) case elements, and c) adverbial modifiers) can become large. To build a manageable controlled language, it would be useful to first focus on a specific text type and examine the growth of coverage according to the lexicon size (Miyata/Sugino 2020). Although our controlled language is based on Japanese, most of the components can also be defined in other languages. Nevertheless, we should carefully reconsider the design of controlled language components when another language is targeted.

We also introduced our prototype of the writing support tool that partly implements the controlled language components. At this stage, the tool purely exploits the constructed controlled language. A wide variety of writing support tools, or augmented writing tools, have been developed to date (Du et al. 2022; Simonsen 2020; Wanner/Verlinde/Alonso Ramos 2013; Yen et al. 2015). To improve the functionality and interface of our tool, it would be effective to utilise technologies found in these existing tools. While our tool is currently intended for post hoc checking scenarios, providing diagnostic functions, more pre-emptive solutions might be useful, such as suggesting subsequent words (e. g., Chen et al. 2019). In future work, we will fully develop the tool and evaluate its usability in practical work scenarios.

References

ASD (2021): ASD Simplified Technical English. Specification ASD-STE100, Issue 8. http://www.asd-ste100.org (last access: 19-03-2022).

Bernth, A./Gdaniec, C. (2001): MTranslatability. In: Machine Translation 16 (3), pp. 175–218.

Chen, M. X./Lee, B. N./Bansal, G./Cao, Y./Zhang, S./Lu, J./Tsay, J./Wang, Y./Dai, A. M./Chen, Z./Sohn, T./Wu, Y. (2019): Gmail smart compose: Real-time assisted writing. In: Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, 4–8 August 2019, Anchorage, Alaska, pp. 2287–2295.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Devlin, J./Chang, M.-W./Lee, K./Toutanova, K. (2019): BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2–7 June 2019, Minneapolis, Minnesota, pp. 4171–4186.

Du, W./Kim, Z. M./Raheja, V./Kumar, D./Kang, D. (2022): Read, revise, repeat: A system demonstration for human-in-the-loop iterative text revision. In: Proceedings of the 1st Workshop on Intelligent and Interactive Writing Assistants, 26 May 2022, Dublin, Ireland, pp. 96–108.

Ishiguro, K. (2016): Setsuzokushi no gijutsu [Techniques of Conjunctions]. Jitsumukyoiku-shuppan. (in Japanese).

Japan Technical Communicators Association (2016): Nihongo sutairu gaido [Style Guide for Japanese Documents] (3rd ed.). Japan Technical Communicators Association. (in Japanese).

Japio (2018): Tokkyo raitingu manyuaru: sangyo nihogo [Patent writing manual: Technical Japanese] (2nd ed.). Japan Patent Information Organization. (in Japanese).

Kato, T./Miyata, R/Tatsumi, M./Sato, S. (2020a): Designing and implementing a support system for simplifying expository text on Japanese Cultural Assets. In: Proceedings of the 34th Annual Conference of the Japanese Society for Artificial Intelligence, 9–12 June 2020, Online, pp. 1–4. (in Japanese).

Kato, T./Miyata, R./Sato, S. (2020b): BERT-based simplification of Japanese sentence-ending predicates in descriptive text. In: Proceedings of the 13th International Conference on Natural Language Generation, 12–15 December 2020, Online, pp. 242–251.

Kawahara, D./Kurohashi, S. (2006): A fully-lexicalized probabilistic model for Japanese syntactic and case structure analysis. In: Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics, 4–9 June 2006, New York, pp. 176–183.

Kittredge, R. (2003): Sublanguages and controlled languages. In: Mitkov, R. (ed.): Oxford Handbook of Computational Linguistics, Oxford University Press, pp. 430–437.

Kuhn, T. (2014): A survey and classification of controlled natural languages. In: Computational Linguistics, 40 (1), pp. 121–170.

Mitamura, T./Baker, K./Nyberg, E./Svoboda, D. (2003): Diagnostics for interactive controlled language checking. In: Proceedings of the Joint Conference Combining the 8th International Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop, May 2003, Dublin, Ireland, pp. 237–244.

Miyata, R./Hartley, A./Paris, C./Kageura, K. (2016): Evaluating and implementing a controlled language checker. In: Proceedings of the 6th International Workshop on Controlled Language Applications, 28 May 2016, Portorož, Slovenia, pp. 30–35.

Miyata, R./Sugino, H. (2020): Building a controlled lexicon for authoring automotive technical documents. In: Gavriilidou, Z./Mitits, L./Kiosses, S. (eds.): Proceedings of XIX EURALEX Congress: Lexicography for Inclusion (Volume 1), 7–9 September 2021, Online, pp.171–180.

Nyberg, E./Mitamura, T./Huijsen, W.-O. (2003): Controlled language for authoring and translation. In: Somers, H. (ed.): Computers and the translator.Amsterdam, pp. 245–281.

Peters, M. E./Neumann, M./Iyyer, M./Gardner, M./Clark, C./Lee, K./Zettlemoyer, L. (2018): Deep contextualized word representations. In: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 1–6 June 2018, New Orleans, Louisiana, pp. 2227–2237.

Sano, M./Sato, S./Miyata, R. (2020): Detection of functional expressions in Japanese sentence-ending predicative phrases and its application to estimation of rhetorical relation between sentences. In:

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Proceedings of the 26th Annual Meeting of the Association for Natural Language Processing, 17–20 March 2020, Online, pp. 1483–1486. (in Japanese).

Sato, S. (2020): HaoriBricks3: A domain-specific language for Japanese sentence composition. In: Journal of Natural Language Processing, 27 (2), pp. 411–444. (in Japanese).

Simonsen, H. K. (2020): Augmented writing and lexicography: A symbiotic relationship? In: Gavriilidou, Z./Mitits, L./Kiosses, S. (eds.): Proceedings of XIX EURALEX Congress: Lexicography for Inclusion (Volume 1), 7–9 September 2021, Online, pp. 509–514.

Teramura, H. (1975–78): Rentaishushoku no shintakusu to imi 1–4 [Syntax and meaning of attributive modification 1–4]. In: Teramura, H. (1992): Teramura Hideo ronbunshuu 1 [Collection of Papers by Hideo Teramura 1]. Kuroshio Shuppan. (in Japanese).

Tolmachev, A./Kawahara, D./Kurohashi, S. (2018): Juman++: A morphological analysis toolkit for scriptio continua. Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, System Demonstrations, 2–4 November 2018, Brussels, Belgium, pp. 54–59.

Wanner, L./Verlinde, S./Alonso Ramos, M. (2013): Writing assistants and automatic lexical error correction: Word combinatorics. In: Kosem, I./Kallas, J./Gantar, P./Krek, S./Langemets, M./Tuulik, M. (eds.): Electronic lexicography in the 21st century: Thinking outside the paper. Proceedings of the eLex 2013 Conference, 17–19 October 2013, Tallinn, Estonia, pp. 472–487.

Warburton, K. (2014): Developing lexical resources for controlled authoring purposes. In: Isahara, H./Choi, K.-S./Lee, S./Nam, S. (eds.): Proceedings of LREC 2014 Workshop: Controlled Natural Language Simplifying Language Use, 27 May 2014, Reykjavik, pp. 90–103.

Warburton, K. (2021): The Corporate Terminologist. Amsterdam/Philadelphia.

Yen, T.-H./Wu, J.-C./Chang, J./Boisson, J./Chang, J. (2015): WriteAhead: Mining grammar patterns in corpora for assisted writing. In: Proceedings of ACL-IJCNLP 2015 System Demonstrations, 26–31 July 2015, Beijing, China, pp. 139–144.

Contact information

Rei Miyata
Nagoya University
miyata.rei.f2@f.mail.nagoya-u.ac.jp

Acknowledgements

This work was supported by JSPS KAKENHI Grant Numbers 19K20628 and 19H05660, and by the the Japan Science and Technology Agency (JST)-Mirai Program (JPMJMI19G8). The automobile manuals used in this study were provided by Toyota Motor Corporation.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Alberto Simões/Ana Salgado

SMART DICTIONARY EDITING WITH LeXmart

Abstract Given the relevance of interoperability, born-digital lexicographic resources as well as legacy retro-digitised dictionaries have been using structured formats to encode their data, following guidelines such as the Text Encoding Initiative or the newest TEI Lex-0. While this new standard is being defined in a stricter approach than the original TEI dictionary schema, its reuse of element names for several types of annotation as well as the highly detailed structure makes it difficult for lexicographers to efficiently edit resources and focus on the real content. In this paper, we present the approach designed within LeXmart to facilitate the editing of TEI Lex-0 encoded resources, guaranteeing consistency through all editing processes.

Keywords Dictionary encoding; Text Encoding Initiative; dictionary editing system

1. Introduction

In the last few years, many scholar projects have been encoding and placing dictionaries online, involving a wide variety of born-digital and retro-digitised lexicographic resources. Conceiving these types of lexicographic resources increasingly requires the application of adapted standards and formats capable of guaranteeing the availability of structured data and ensuring interoperability between systems, especially when the lexicographic production scenario is very heterogeneous due to its nature, form, and content. There are several types of dictionaries, in several languages, with disparate structures and different functions, purposes and users. Many of these dictionaries adopt a hierarchical data structure representation, mainly based on eXtensible Markup Language (XML).

The application of standard formats implies two different aspects: modelling and encoding. Modelling refers to the creation of an abstract model, that can account for all the lexical data and their components (Godfrey-Smith 2009). Encoding refers to the process of expressing the specific lexical data using a predefined data format. Essentially, modelling is a design task, and encoding is an implementation task. These are crucial issues for lexicography to ensure interoperability between the software components of heterogeneous lexicographic resources (Romary/Wegstein 2012).

Dictionaries are modelled and encoded in multiple diverse formats, being the information organised and stored in files of different nature. For dictionary encoding using XML, there are different structured data formats, such as the Dictionary module from the Text Encoding Initiative (TEI Consortium 2021), the XML Dictionary eXchange Format (Snegov/Soshinskiy 2019) or even the OntoLex-Lemon (McCrae et al. 2017).

Currently TEI is the dominant format for several lexicographic projects, such as BASNUM,[1] Nénufar,[2] ARTFL,[3] VICAV[4] or Berlin-Brandenburg Academy of Sciences for digitising and transcribing legacy dictionaries.[5] From the very beginning, the TEI Guidelines have a mod-

[1] https://anr.fr/Project-ANR-18-CE38-0003.

[2] http://nenufar.huma-num.fr/.

[3] https://artfl-project.uchicago.edu/.

[4] https://vicav.acdh.oeaw.ac.at/.

[5] https://gitlab.com/xlhrld/retro-dict.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

ule explicitly focused on the encoding of dictionaries. However, this module is criticised regarding its extreme flexibility, i. e., the existence of multiple possibilities to encode similar structures that affect the interoperability of the encoded formats. In some cases, TEI makes no binding requirements for the possible values since there are many possibilities across different projects. In each lexicographic project, it is likely that standardising an agreed set of values will be very helpful. In this sense, it is better to customise or change the schema by providing more restrictions. This explains why, for example, there is the need to restrict the scope of usage information (Salgado et al. 2019). Interestingly, this flexibility is also the characteristic that justifies its wide adoption.

To reduce this freedom and define a specific format for dictionaries, the TEI Lex-0 initiative (Bański/Bowers/Erjavec 2017; Romary/Tasovac 2018; Tasovac/Romary 2018), a stricter version of the TEI schema, has been promoted to reduce the encoding options. In the context of the ELEXIS project[6], TEI Lex-0 has been adopted as one of the baseline formats (McCrae et al. 2019). Other projects, such as our case-study, *Dicionário da Língua Portuguesa*[7] (ACL 2021), are also using TEI Lex-0 as an encoding format (Salgado et al. 2019).

Within the development of LeXmart[8] (Simões/Salgado/Costa 2021), this schema was also adopted. Nevertheless, while some of the verbosity of the schema helps in the automatic processing, sometimes the proposed encoding can be hard for lexicographers (examples of these structures will be discussed through this paper). To make it easier for the continuous editing of large lexical resources, LeXmart uses a layer of macros to hide this complex structure. This feature also allows the consistency of annotation, as we will discuss shortly.

While our proposal is specifically tailored for TEI Lex-0, the idea behind our approach can be replicated to other formats.

In the next section, we will provide a brief discussion of the TEI data format, as well as dictionary writing systems (DWS). We will focus on the management of the formats, and how the tools deal with them. In section 3, we will describe examples of the complex structures we identified in TEI Lex-0 and present how they were hidden using our macro approach. We will also explain how these structures are taken into consideration for the online rendering of the dictionary. In section 4, the technical details of this layer implementation will be detailed. We conclude with some insights on our approach in section 5.

2. Lexicographic Encoding and Editing

Regarding encoding, the approaches used for lexicographic content follow the same ideas that are used for encoding other types of resources: visual or semantic encoding. A dictionary can be seen as a textual artefact with its own specific publishing history and its own verbal expression and visual arrangement of the linguistic content contained within it or we can instead prioritise the linguistic content, ignoring how it is presented and the exact sequence of words used in, for example, the definitions of articles. There is also a third (poten-

6 https://cordis.europa.eu/project/id/731015.

7 Currently, it is being prepared under the Instituto de Lexicologia e Lexicografia da Língua Portuguesa's supervision in collaboration with researchers and invited collaborators. This project is supported by a small annual Community Support Fund Portuguese National Fund (Fundo de Apoio à Comunidade – FAC) through the Fundação para a Ciência e a Tecnologia (FCT).

8 https://lexmart.eu/.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

tially more verbose) approach, that is, to do both simultaneously and make sure both kinds of information are aligned.

These views are defined as follows: the typographical view aims to mirror the physical structure of a document using elements from the core module. It concerns the layout of individual pages and is mostly used on retro-digitisation projects, where the aspect used to present the information should be kept. These TEI elements can be used to encode the page layout, column and line breaks and highlighted words. Some elements can also be typed to provide more precision on how they are typographically presented in the original printed document. The second level of encoding deals with the semantic and logical function of text structures and is concerned with the conceptual or linguistic content of a dictionary as a whole as well as its individual entries.

While some formats are focused only on content encoding, other are more versatile, allowing these two different encoded approaches in the same document as referred above. As an example, the XDXF format is more focused on content, while TEI is versatile enough to cover both aspects in the same document.

To overcome these constraints, the TEI Lex-0 main goal is to take advantage of the TEI work but introduce stricter rules on how elements can be used. This new paradigm is quite important, as it enforces reusability and interoperability between different systems, and allow easier manipulation of the resources by computational tools. Nevertheless, it is more verbose and, while reusing some element names for different information types, it is more error-prone when the resources editing is made manually.

In this regard, it gets relevant to use a proper editor that is aware of the specific structure of a dictionary and allows the lexicographer to focus on its real work and not in the details of the annotation schema. This means that, while possible, it is better to avoid that a lexicographer edits each dictionary entry directly in a generic XML editor.

DWS are available for some time (Abel 2012), but mostly as in-house tools developed by publishing companies. Recently, free and open-source tools have been made available, such as Lexonomy[9] (Měchura 2017) and LeXmart (Simões/Salgado/Costa 2021).

Concerning DWS, the first solutions were based on forms. The use of forms allows the editing of new lexicographic articles in a clean environment. However, it is not clear how each entry maps to a specific XML element, hiding the true structure of the document. For example, in the case of Lexonomy and LeXmart, the editing is based on a specific XML schema. While for Lexonomy the user can configure their own schema, LeXmart focuses on guaranteeing interoperability and therefore enforces the use of TEI Lex-0.

Both editors suffer from the same problem: the verbosity of the used schema, as the user is presented with the complete XML structure of each entry. This paper proposes a macro layer to simplify the editing of complex XML structures. This idea was implemented on top of LeXmart.

3. TEI Lex-0 Common Patterns

TEI Lex-0 schema has a specific set of common patterns when applying its encoding rules into a specific dictionary. As a starting point, and as a motivational example, consider the

[9] https://www.lexonomy.eu/.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

case of synonyms encoding. According to the TEI Lex-0 schema, these lexicographic components should be encoded as the following structure:

```
<xr type="synonymy">
   <ref type="entry">word</ref>
</xr>
```

While this structure can be a little different in specific situations (for instance, when the synonym is accompanied by a geographic label that identifies the place or region where a lexical unit is mainly used), it will be used and reused for every synonym. This solution encodes the synonym as a cross-reference inside the dictionary and reuses the `<xr>` and `<ref>` elements, making it clear the internal reference between entries and, at the same time, reusing elements that are useful in other contexts of the encoding process. Nevertheless, presenting this XML directly to the lexicographer can be confusing and overwhelming as, in some way, it distracts the lexicographers from their real work of linguistic analysis.

Bearing this in mind, we considered the possibility of using special elements, working as macros. By macro, we mean custom XML elements that are expanded to and from a more complex structure of elements. This allows the replacement of a common structure, like the one shown above for encoding synonyms, by a non-standard element that hides the original element tree. For instance, the use of an artificial element, named `<syn>`, especially used for this purpose:

```
<syn>word</syn>
```

This approach allows the user to understand the proper structure of the entry and edit all the lexicographic content faster and with less visual clutter. Thus, the lexicographer will just see this element in the DWS. The tool performs the necessary steps to guarantee that the document structure will automatically be converted to and from the original schema. LeXmart will also allow the lexicographer to choose between the original XML structure or, instead, edit the simplified version.

The next subsections discuss this approach, detecting such patterns, and propose custom and simplified custom elements, that are used only for presenting the information to lexicographers. We discuss each simplified structure, exemplifying with lexicographic articles from the new *Dicionário da Língua Portuguesa* (DLP) (ACL 2021). This lexicographic work is a retro-digitised dictionary (Simões/Almeida/Salgado 2016) whose starting point was the *Dicionário da Língua Portuguesa Contemporânea* (ACL 2001), last published in 2001.

3.1 Synonyms and antonyms

Just as mentioned above, synonyms and antonyms are encoded as cross-references to their own entries in the dictionary. While making sense, the structure can be overwhelming. We suggest the use of the `<syn>` and `<ant>` elements as a shortcut to include this kind of reference. The main issue arises when there is further information associated with these elements as, for instance, a specific geographic area where that synonym is used, or a specific domain of knowledge where that synonym is mainly used.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

As an illustrative example, consider the sense of *casmurro*[2] [pigheaded] entry with the following synonym:

```
<xr type="synonymy">
   <ref type="entry">
      <usg type="socioCultural">Inf.</usg>
      cabeçudo
   </ref>
</xr>
```

This example includes a usage label (*socioCultural* type),[10] which restricts the scope of the synonym in contexts of informality. When applying the `<syn>` macro, this part of the entry gets simplified to:

```
<syn>
   <usg type="socioCultural">Inf.</usg>
   cabeçudo
</syn>
```

Therefore, the usage information is not lost. This process is applied in both directions, thus generating the original markup when required.

3.2 Etymology Cross-references

When encoding etymology and data components, including foreign language words (for instance, the origin Latin word), TEI Lex-0 allows the lexicographer to include a full entry registering that word sense, directly in the etymological section. While powerful, this also requires a complex structure. Consider the following example from the *era* [era] entry:

```
<etym>
   Do latim
   <cit type="etymon">
      <form><orth xml:lang="la">aera</orth></form>
   </cit>
</etym>
```

[10] See 8.2. Types of usage: https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html#index.xml-body.1_div.8_div.2.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

This structure emphasizes that the Latin word is a citation, from an etymological point of view, which form is, in Latin, *aera*. While it is possible that in some situations this structure can be useful to add further details, that will rarely be the case. In a dictionary for a Romance language such Portuguese, where most words are derived from Latin, the editing of this information gets repetitive and time-consuming. Thus, we propose the use of the `<etymon>` element, which works as a macro for etymological citations:

```
<etym>
   Do latim
   <etymon xml:lang="la">aera</etymon>
</etym>
```

This kind of annotation is not only smaller but can also be presented more cleanly in the online DWS.

3.3 Examples

In TEI Lex-0 the `<cit>` and `<quote>` elements are used together for different purposes, namely for the inclusions of bibliographic examples (illustrative quotations extracted from corpora obtained from known authors) and usage examples (examples made up by the lexicographer). The distinction of the two types is obtained by attributes added in the `<cit>` element. A common example in DLP (from the entry *casmurro*[2] [pigheaded]) is codified as[11]:

```
<cit type="example">
   <quote type="example">
      Por ser tão casmurro, perdeu a oportunidade de
      fazer um bom negócio.
   </quote>
</cit>
```

For the bibliographic citations, they are encoded using these same two elements, but with different attributes, and with extra bibliographic information. Consider the following example from the *amor* [love] entry:

```
<cit type="example">
   <quote>
      A noite é para o amor e o amor para Guma é Lívia.
```

[11] The double use of the @type attribute with the value of "example" is a choice of the authors for codifying lexicographic examples and differentiate them from bibliographic examples. This use is not generalized, as this distinction can be done by checking the existence of the bibliographic information inside the citation. Nevertheless, this decision was taken to make formatting through Cascading Style Sheets easier.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

```
        Ele não quer amor de aventuras, amor de acaso.
    </quote>
    <bibl>
        <author>J. Amado</author>
        <title>Mar</title>
    </bibl>
</cit>
```

This type of example is more complex given the introduction of bibliographic information. Thus, we decided not to simplify it. But the clear distinction between both is important and must be presented clearly for the lexicographer (and end-user). Bearing this in mind, a simpler version of the lexicographic example was developed. The example presented above for a usage example is presented in the LeXmart as:

```
<example>
    Por ser tão casmurro, perdeu a oportunidade de
    fazer um bom negócio.
</example>
```

This distinction, together with the fact that the editor does not allow the inclusion of bibliographic information in an `<example>` tag, allows the lexicographer to quickly identify the type of quote.

3.4 Polylexical units

Many of the lexicographic articles from DLP include polylexical units, i.e., "a stable and recurrent sequence of lexemes that are perceived as an independent lexical unit by the speakers of a language" (Tasovac/Salgado/Costa 2020, p. 29), commonly called multiword expressions, collocations, lexical combinations, or even "*co-ocorrente privilegiado*"[12] [privileged co-occurrent]. Tasovac/Salgado/Costa (2020) argue that the encoding of polylexical units in dictionaries is a topic that has not been covered adequately and in sufficient depth by the TEI regarding the formal representation of polylexical units as they appear on the page of a single dictionary. The authors, for the case of privileged co-occurrent, recommend encoding this last type of polylexical units as `<form>` elements as they are presented to the end-user as a sequence of forms. However, and as this amendment has not yet been implemented, the DLP still maintains these sequences as a special kind of example as we explain further.

[12] Privileged co-occurrent is a dependency relationship ("uma relação de dependência") which occurs between full words ("palavras plenas") such as nouns, adjectives, verbs and adverbs and other words in the construction of sentences ("na construção das frases") (ACL 2001, p. XXI).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

descalçar [diʃkałs'ar]

verbo

1. tirar (o que calça os pés, mãos ou pernas)

ANTÓNIMOS calçar ; enfiar ; pôr

descalçar as botas, as luvas, as meias

descalçar os sapatos

Fig. 1: Snippet from the *descalçar* entry from DLP

Figure 1 shows the first sense of the *descalçar* [take of the shoes] lemma in the DLP. The two last lines, "*descalçar as botas, as luvas, as meias*" [to remove one's boots, one's gloves, one's socks] and "*descalçar os sapatos*" [to remove one's shoes], illustrate this type of polylexical units, that function as "*blocos semântica e sintaticamente afins*" [semantically and syntactically related blocks] (ACL 2001, p. XXI). In other words, the aim is to show that the lemma *descalçar* occurs frequently with the given nouns (boots, gloves, socks, shoes).

These polylexical units are encoded as a citation, just like the two types of examples presented before, but they need to be presented in a different way, both to the end-user and to the lexicographer. We applied a new macro for this structure. Consider the following fragment from Fig. 1 entry:

```
<cit type="example">
   <quote type="collocation">
      <hi>descalçar</hi> as botas, as luvas, as meias
   </quote>
</cit>
<cit type="example">
   <quote type="collocation">
      <hi>descalçar</hi> os sapatos
   </quote>
</cit>
```

Just like in previous situations, we decided to create a simple macro `<collocation>` to hide this structure and differentiate common bibliographic examples from this specific type:

```
<collocation>
   <hi>descalçar</hi> as botas, as luvas, as meias
</collocation>
<collocation>
   <hi>descalçar</hi> os sapatos
</collocation>
```


This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

4. LeXmart implementation details

LeXmart is built on top of eXist-DB[13] as main backend. This document-oriented noSQL database is built with support for W3C technologies[14] like XQuery, XPath, XForms and XSLT. This allows the usage of XSL transformations to perform the conversion between the official TEI Lex-0 format and the simplified version described earlier.

The conversion process is composed of two XSLT files, one to simplify the notation, and another one to restore the correctness of the document. These stylesheets are prepared to do not perform any change when there is a structure that does not fit exactly on the pattern that was defined. This way, it is possible to ensure that there will be no loss of information during the conversion process.

To keep this pair of stylesheets working correctly, their composition needs to have the same behaviour as the mathematical identity function: return the original document.

Figure 2 shows the portion of the stylesheet that is responsible to convert a citation inside the etymology element to a `<etymon>` element.

```
<xsl:template match="tei:cit[@type='etymon']">
  <xsl:choose>
    <xsl:when test="./tei:form/tei:orth">
      <etymon>
        <xsl:if test="./tei:form/tei:orth/@xml:lang">
          <xsl:attribute name="xml:lang">
            <xsl:value-of
                    select="./tei:form/tei:orth/@xml:lang"/>
          </xsl:attribute>
        </xsl:if>
        <xsl:apply-templates
                    select="./tei:form/tei:orth/node()"/>
      </etymon>
    </xsl:when>
    <xsl:otherwise><xsl:apply-templates/></xsl:otherwise>
  </xsl:choose>
</xsl:template>
```

Fig. 2: Extract from the XSLT template to perform the conversion of citations inside the etymology element

[13] https://exist-db.org/

[14] https://www.w3.org/

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Meanwhile, the inverse process is obtained applying the stylesheet presented in Figure 3.

```
<xsl:template match="tei:etymon">
  <cit type="etymon">
    <form>
      <orth>
        <xsl:if test="@xml:lang">
          <xsl:attribute name="xml:lang">
            <xsl:value-of select="@xml:lang"/>
          </xsl:attribute>
        </xsl:if>
        <xsl:apply-templates/>
      </orth>
    </form>
  </cit>
</xsl:template>
```

Fig. 3: Extract from XSLT template to reverse the conversion of the `<etymon>` element inside the etymology element

These transformations can be performed on server-side, applying the stylesheets when the entries are fetched or saved in the database, or performed directly on client side, using the lexicographer's browser.

5. Final remarks

The use of XML to encode any type of content has always created discussion regarding its verbosity. The truth is that, adding elements to a digital content, and properly annotating it can duplicate the size of the document. That is a reason why XML as a serializing format is being less and less used (Fonseca/Simões 2007), in comparison with other formats such as JSON (JavaScript Object Notation) or YAML (Yet Another Markup Language).

Nevertheless, for digital humanities, the requirement of annotating texts at different levels, and not just as a structural schema, requires the use of *mixed content elements* (elements that can contain text and other elements as direct children), which are only possible with XML. But to be effective, the use of XML needs to be complemented with tools that allow content authors to quickly annotate their information.

In this paper, we proposed a solution based on macros to simplify some structures of TEI Lex-0 for editing purposes, making the user interface simpler to the lexicographer. The process is transparent to the lexicographers, in the sense that they do not need to be aware of the conversion process to properly use the tool.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

The irreversible transition to the digital environment has imposed on lexicography (and the humanities and social sciences in general) the challenge of adopting new methods concerning the traditional ones. It is important to highlight that we are still in a transition phase, where lexicographers, who worked for many years on printed dictionaries, are making efforts to embrace the digital environment. The traditional lexicographer no longer exists; today, any e-lexicographer must be a digital humanist, being, for example, also an encoder. This is our real concern: to help lexicographers without encoding experience to be able to edit lexical resources with more confidence and to allow them to dedicate themselves to what is really important, the lexicographic work *per se*.

References

Abel, A. (2012): Dictionary writing systems and beyond. In: Granger, S./Paquot, M. (eds.): Electronic lexicography. Oxford, pp. 83–106.

ACL (2001): Dicionário da Língua Portuguesa Contemporânea. Malaca Casteleiro, J. (Coord.). Lisbon: Academia das Ciências de Lisboa and Editorial Verbo.

ACL (2021): Dicionário da Língua Portuguesa. Salgado, A. (Coord.) [New digital edition under revision]. Lisbon.

Bański, P./Bowers, J./Erjavec, T. (2017): TEI-Lex0 Guidelines for the Encoding of Dictionary Information on Written and Spoken Forms. In: eLex 2017: Lexicography from Scratch. Leiden, pp. 485–494.

Fonseca, R./Simões, A. (2007): Alternativas ao XML: YAML e JSON. In: XATA 2007- 5ª Conferência Nacional em XML, Aplicações e Tecnologias Associadas, pp. 33–46.

Godfrey-Smith, P. (2009): Models and fictions in science. In: Philosophical Studies, 143, pp. 101–116. http://doi.org/10.1007/s11098-008-9313-2.

McCrae, J. P. et al. (2017): The OntoLex-Lemon Model: development and applications. In: eLex 2017: Lexicography from Scratch. Leiden, pp. 587–597.

McCrae, J P. et al. (2019): The ELEXIS Interface for Interoperable Lexical Resources. In: eLex 2019 – Electronic Lexicography in the 21st Xentury, pp. 642–659.

Měchura, M. (2017): Introducing lexonomy: an open-source dictionary writing and publishing system. In: eLex 2017 – Lexicography from Scratch, pp. 662–679.

Romary, L./Tasovac, T. (2018): TEI Lex-0: a target format for TEI-Encoded dictionaries and lexical resources. In: 8th Conference of Japanese Association for Digital Humanities, pp. 274–275.

Romary, L./Wegstein, W. (2012): Consistent modeling of heterogeneous lexical structures. In: Journal of the Text Encoding Initiative 3. http://doi.org/https://doi.org/10.4000/jtei.540.

Salgado, A. et al. (2019): TEI Lex-0 in action: improving the encoding of the dictionary of the Academia das Ciências de Lisboa. In: eLex 2019 – Electronic Lexicography in the 21st Century, pp. 417–433.

Simões, A./Almeida, J. J./Salgado, A. (2016): Building a dictionary using XML technology. In: 5th Symposium on Languages, Applications and Technologies (SLATE'16), pp. 14:1–14:8. http://doi.org/10.4230/OASICS.SLATE.2016.14.

Simões, A./Salgado, A./Costa, R. (2021): LeXmart: a platform designed with lexicographical data in mind. In: eLex 2021 – Post Editing Lexicography, pp. 529–541.

Snegov, S./Soshinskiy, L. (2019): Why is XDXF better than other dictionary formats? GitHub. https://github.com/soshial/xdxf_makedict (last access: 22-03-2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Tasovac, T./Romary, L. (2018): TEI Lex-0: a baseline encoding for lexicographic data. Version 0.8.5. DARIAH Working Group on Lexical Resources. https://dariah-eric.github.io/lexicalresources/pages/TEILex0/TEILex0.html# (last access: 22-03-2022).

Tasovac, T./Salgado, A./Costa, R. (2020): Encoding polylexical units with TEI Lex-0: a case study. In: Slovenščina 2.0, 2, pp. 28–57.

TEI Consortium (2021): TEI P5: guidelines for electronic text encoding and interchange. TEI Consortium. http://www.tei-c.org/Guidelines/P5/ (last access: 22-03-2022).

Contact information

Alberto Simões
2Ai – School of Technology, IPCA, Barcelos, Portugal
asimoes@ipca.pt

Ana Salgado
NOVA CLUNL, Centro de Linguística da Universidade NOVA de Lisboa, Portugal/Academia das Ciências de Lisboa, Portugal
ana.salgado@fcsh.unl.pt

Acknowledgements

This paper was partially funded by Portuguese national funds (PIDDAC), through the FCT – Fundação para a Ciência e Tecnologia and FCT/MCTES under the scope of the projects UIDB/05549/2020 and UID/LIN/03213/2020, and by the FCT – Fundação para a Ciência e Tecnologia as part of the project Centro de Linguística da Universidade NOVA de Lisboa – UID/LIN/03213/2020.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Design and Publication of Dictionaries

Zita Hollós

CROSS-MEDIA-PUBLISHING IN DER KORPUSGESTÜTZTEN LERNERLEXIKOGRAPHIE

Entstehung eines Lernerwörterbuchportals DaF

Abstract This paper gives an insight into a cross-media publishing process on different stages: from a printed bilingual syntagmatic dictionary for GFL to an online learner's dictionary of German collocations to a German learner's dictionary portal. On the basis of an sql database specially developed for a corpus-guided dictionary of German collocations, the bilingual syntagmatic learner's dictionary KolleX was published in 2014. The first part of the article describes this lexicographic process, focusing the most relevant aspects of the dictionary concept, e. g. dictionary type, subject matter, corpus-guided data selection and microstructure. The second part introduces the first online version of KolleX from 2016 and the profound changes in the editing system – from a desktop version (2005) to a web-based editing system (2016) –, which resulted successively in a prototype of a German learner's dictionary portal, called E-KolleX DaF (2018–). Focusing on the aspects of dynamism and integration of different resources from a learner's perspective the paper shows the innovative features of this new online reference work. The contribution presents the solutions for the integration of new datatypes in the database of KolleX and the linking to different data in German monolingual dictionary platforms. The paper outlines the web design, functioning and technical improvements of E-KolleX DaF. The conclusions provide an outlook to the forthcoming challenges.

Keywords Wörterbuchportal; Kollokationen; Lernerwörterbuch; DaF; Datenbank; Webdesign

1. Etappe I: KolleX DaF, das Printwörterbuch (2004–2014)

1.1 Datenbankbasiertes und korpusgestütztes Wörterbuchprojekt für Kollokationen

Der erste Band des Nachschlagewerks „Wörterbuch zur Lexikographie und Wörterbuchforschung" (WLWF) definiert *Cross-Media-Publishing* folgendermaßen: „Publikationsprozess, der darauf hin konzipiert ist, aus ein- und derselben Datenbasis Produkte für beliebige Publikationsmedien zu erzeugen" (WLWF 2010, S. 751). Im vorliegenden Beitrag wird dementsprechend das Ziel verfolgt, die Ergebnisse eines solchen Prozesses zu präsentieren: wie auf der Grundlage einer Datenbank für deutsch-ungarische Kollokationen ein syntagmatisches Printwörterbuch DaF publiziert wurde und daraus wenig später ein Internetwörterbuch für Kollokationen bereits für ein breites Lernerpublikum entstand. Gleichzeitig wird der Artikel einen ersten Überblick über die Etappen eines datenbankbasierten lexikographischen Prozesses mit geplanter Print- und dann digitaler Publikation geben. Wie letztendlich ein Prototyp eines Wörterbuchportals DaF sukzessive auf dieser Datengrundlage entwickelt wurde, konnte bisher umfassend nicht erörtert werden. Der vorliegende Beitrag will auch diese Lücke schließen.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

1.2 Zweisprachiges syntagmatisches Printwörterbuch DaF

Moderne Korpus- und Sprachtechnologie sowie Verwaltung und Modellierung von Daten in einer Datenbank sind heutzutage beim Aufbau verschiedener Sprachressourcen ebenso wenig wegzudenken wie die Datenpräsentation, also die Planung einer dynamischen Benutzeroberfläche und eines attraktiven Webdesigns (vgl. Klosa/Müller-Spitzer 2016).

Als 2005 das Forschungs- und Wörterbuchprojekt SZÓKAPTÁR/KOLLEX (im Weiteren KolleX) mit dem Ziel der Erstellung eines korpus- und datenbankbasierten zweisprachigen Kollokationswörterbuchs für Deutschlerner gestartet wurde, stand die Korpus- und Sprachtechnologie für das Deutsche noch relativ am Anfang. Die Realisierung datenbankbasierter elektronischer Wörterbücher – die im Rahmen einsprachiger Ressourcen wie *elexiko* bereits erprobt wurde (vgl. Hass 2005) – war ein komplett neues Terrain für die zweisprachige Speziallexikographie, insbesondere für die Kollokationslexikographie für Sprachlerner.

In diesem Kapitel werden die wichtigsten Phasen eines speziallexikographischen Projekts vorgestellt, dessen Ergebnisse sowohl als Printpublikation (Hollós 2014) als auch teilweise in digitaler Form bereits im Jahre 2016 vorlagen. Seitdem entsteht sukzessive ein Wörterbuchportal für ungarischsprachige Deutschlerner.[1]

1.2.1 Forschungsprojekt für ein syntagmatisches Lernerwörterbuch

Die theoretischen Vorarbeiten lagen bereits in Form einer Lernerwörterbuchkonzeption (vgl. Hollós 2004) vor. Ohne diese Planungsphase (2000–2003) wäre die Phase der Datenmodellierung und -speicherung in einer sql-Datenbank sowie die Erstellung eines umfangreichen, zweisprachigen, korpusgestützten Lernerwörterbuchs der Kollokationen für Deutschlerner (2004–2014) nicht möglich gewesen.

Für das durch den *Ungarischen Nationalfonds für Wissenschaft und Forschung* (OTKA) von 2005–2008 finanzierte Forschungsprojekt wurde eigens ein Datenbank- und Redaktionssystem mit vielen modernen Funktionen für das geplante Spezialwörterbuch entwickelt. Im Rahmen der lexikographischen Werkstatt der Universtität wurden ein Informatiker, im Durchschnitt drei stundentische Mitarbeiter und zwei deutsche Muttersprachler als Mitarbeiter und Lektoren beauftragt. Die Korpusdaten für das Kollokationswörterbuch sind der Zusammenarbeit mit dem Leipziger Projekt *Deutscher Wortschatz* (DW) zu verdanken.

Als Endprodukt ist ein zweisprachiges, syntagmatisches Lernerwörterbuch für Deutschlerner der Mittel- und Oberstufe entstanden, das insgesamt mehr als 61.000 Kollokationen, Wortverbindungen, Kombinationen mit Valenzangaben enthält, von denen mehr als 48.000 auch ins Ungarische übersetzt wurden. Im KolleX sind somit mehr als 10.000 verschiedene Lexeme als Kollokatoren/Kotextparter zu 2.262 Basis-Lemmata verzeichnet. Die exakten Zahlen des zweisprachigen, syntagmatischen Printwörterbuchs SZÓKAPTÁR/KOLLEX präsentiert die Tabelle 1:

[1] E-KolleX (DaF) ist seit 2016 unter der folgenden Internetseite zu finden: www.kollex.hu/szotar.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Basislemmata (im Wörterverzeichnis)	2262
Kollokatoren/Kotextpartner (im Register)	10313
Kollokationen/Wortverbindungen (nach Strukturtypen)	48757
Kombinationen (eigenständiger Strukturtyp für komplexe Kollokationen)	2661
Valenzrealisierungen (bei Verblemmata)	8590
Basen zu den Kollokatoren (bei Adjektivlemmata)[2]	1609
Kollokationen/Kotextpartner insgesamt	**61617**

Zahlen von SZÓKAPTÁR/KOLLEX

Tab. 1: Exakte Zahlen von Lemmata, Kollokationen/Wortverbindungen, Kombinationen u. a.

1.2.2 Wörterbuchtyp, Wörterbuchgegenstand und Datenselektion

Im Folgenden werden ausgewählte Aspekte der Wörterbuchkonzeption, wie der Wörterbuchtyp, der Wörtebuchgegenstand sowie die korpusgestützte Datenselektion (i. S. v. Klosa 2011, S. 112) erörtert.[3]

Die ursprüngliche Wörterbuchcharakteristik (vgl. Hollós 2004, S. 174) musste im Hinblick auf das Ergebnis des publizierten Printwörterbuchs hin modifiziert werden: KOLLEX (2014) ist ein korpusgestütztes und datenbankbasiertes, polyfunktionales, dennoch produktionsbezogenes zweisprachiges syntagmatisches Lernerwörterbuch, dessen primärer Wörterbuchgegenstand die deutschen Kollokationen zu den Autosemantika der Wortliste des Zertifikats Deutsch[4] umfasst.

Im KOLLEX wird mit einem neuentwickelten Kollokationsbegriff gearbeitet. Die fremdsprachendidaktisch motivierte Kollokationsauffassung (i. S. v. Hausmann 1984) wurde im Rahmen eines integrativen Kollokationsbegriffs mit den Ergebnissen der Sprachtechnologie (vgl. Quasthoff 2009) kombiniert.[5] Statistische Kookkurrenzprofile wie auch im DWDS-Wortprofil (vgl. Geyken 2011) liefern große Mengen an sprachlichen Daten für statistisch signifikante Kookkurrenzpartner. Viele sind i. o. S. zur Basis affine Kollokatoren. Ausgehend von der Basis, deren Sememstruktur für alle Wörterbucheinträge als erstes erarbeitet werden musste, können mit (beinahe) muttersprachlicher Kompetenz typische Kollokatoren/Kotextpartner sememspezifisch ausgewählt und übersetzt, bzw. mit stilistisch-pragmatischen Angaben sowie mit morphosyntaktischen Restriktionen versehen werden.

[2] Mit der Umkehrung der gerichteten Kollokationsrelation bei den Lemmazeichentypen Adjektiv und Verb wurden neue Datentypen in den Wörterbuchartikeln etabliert: Bei den adjektivischen Lemmata – in diesem Fall nicht als Basis, sondern als Kollokator verstanden – werden auch Kollokationsbasen, also typische Substantive und Verben aufgelistet. Ebenfalls kann man die substantivischen Valenzrealisierungen als Basen zum Verb (Kollokator) interpretieren.

[3] Aus Platzgründen beschränke ich mich auf eine Zusammenfassung der Ergebnisse des Wörterbuchprojekts zwischen 2004–2014 und verweise auf die Projekthomepage mit einzelnen Publikationen zu spezifischen Aspekten der Wörterbuchkonzeption und -arbeit: www.kollex.hu.

[4] Einzelheiten zur Lemmaselektion sind in Hollós (2004, S. 154–159) zu finden.

[5] Näheres zu diesem Kollokationsbegriff findet man in Hollós (2016). Die Unterscheidung in intra- und interlinguale Kollokationen integriert auch den kontrastiven Aspekt und bildet damit die theoretische Grundlage für das ursprünglich geplante zweite Wörterverzeichnis der intralingualen Kollokationen, deren automatische Generierung die alte Desktop-Version des Programms noch immer ermöglicht. Das Printwörterbuch enthält außerdem 8378 mit dem Symbol „☝“ markierte, interlinguale Kollokationen.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Da zur Zeit der Datenselektion und -aufbereitung das DWDS-Wortprofil noch nicht zur Verfügung stand, wurden die Korpusdaten für das Wörterbuch in Zusammenarbeit mit dem Leipziger Korpus- und Wortschatzprojekt gewonnen.[6] Darüber hinaus wurde die am IDS entwickelte *Kookkurrenzdatenbank* (CCDB) von Belica (1995) genutzt sowie Daten aus den damals modernsten ein- und zweisprachigen Wörterbüchern übernommen. Das syntagmatische Lernerwörterbuch KolleX ist dementsprechend korpusgestützt (i. S. v. Klosa 2011, S. 112)[7] erarbeitet worden, da bei der Selektion und Anordnung der Kollokatoren/Kollokationen korpusgesteuert, bei der Angabe der Valenz und der Auswahl der Valenzrealisierungen bei den Verblemmata korpusbasiert, mit den Daten von VALBU gearbeitet wurde.[8]

1.3 Artikelstruktur im KolleX

Die folgende Abbildung zeigt einen typischen Wörterbuchartikel aus dem KolleX:[9]

E-Mail ['i:meɪl] die / dn/oszt/sv das, der E-Mail/des E-Mails, die E-Mails <fn> **e-mail**
SUBS **eine Menge E-Mails** ☞ rengeteg/egy csomó e-mail
ADJ **dringend** sürgős • csak jel **unerwünscht** kéretlen, nem kívánatos • **formlos** nem hivatalos, kötetlen formájú • **unpersönlich** személytelen • **infiziert** vírusos
VERB kissé vál **senden**[AKK] (el)küld • **bekommen**[AKK] kap • **speichern**[AKK] (el)ment • **löschen**[AKK] töröl • **weiterleiten**[AKK] továbbít • vál **archivieren**[AKK] megőriz, archivál • **markieren**[AKK] *pl. továbbításra, törlésre* kijelöl • **auf eine E. antworten** ☞ válaszol egy e-mailre • **etw**[Akk] **per E. bestellen** ☞ e-mailben rendel (meg) vmit • **etw**[Akk] **per E. verbreiten** ☞ e-mailben terjeszt vmit
KOMB **eine dringend zu beantwortende E.** sürgősen megválaszolandó e-mail

wa_1 zu **E-Mail** im KolleX

Abb. 1: Darstellung der Artikelstruktur zum Lemmazeichen E-Mail, Abkürzungen: dn/oszt/sv = sdt./österr./schweiz., fn = Substantiv, csak jel = nur attributiv, kissé vál = ein wenig gewählt

Im Umtext „Kurze Benutzungshinweise" findet man eine detaillierte Beschreibung der Artikelstruktur, deren Anfang hier wiedergegeben wird:

> Der zentrale Gegenstand des Wörterbuchs sind Kollokationen und typische Wortverbindungen zu den einzelnen Stichwörtern. Sie werden in den Wörter-

[6] Die Analyse erfolgte mit dem Log-Likelihood-Maß von Dunning (vgl. Quasthoff 2009, S. 159).

[7] Annette Klosa (2011, S. 110ff.) verwendet den Terminus „korpusgestützt" oberbegrifflich für „korpusvalidierend" (engl. copus-based) und „korpusgesteuert" (engl. corpus-driven).

[8] Näheres findet man zur korpusgesteuerten und -basierten Datenselektion und -aufbereitung in Bezug auf die Verblemmata in Hollós (2008).

[9] Die Online-Version des Wörterbuchartikels findet man in Abbildung 2.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

> buchartikeln wortartspezifisch, nach den Strukturtypen der Kollokationen angeordnet:
> zu den Substantiven **Substantive, Adjektive** und **Verben**
> zu den Verben **Adverbien**
> zu den Adjektiven (und Adverbien) **Adverbien.**
>
> Teils werden nur die obigen Partner/Kollokatoren (bei Verben oft mit der jeweiligen Valenz) genannt, teils auch das abgekürzte Stichwort mitverzeichnet und damit die ganze Wortverbindung/Kollokation angegeben. […]
>
> Bei allen Lemmazeichentypen kann die Kategorie der **Kombinationen** auftreten. Hier werden größere, schwach oder teil-idiomatische Wortverbindungen verzeichnet. Das abgekürzte Lemma ist bei diesen Kombinationen obligatorischer Teil der Verbindung. (KolleX 2014, S. 963)

Außerdem ist hervorzuheben, dass die Kollokationen/Wortverbindungen mit ihren morphosyntaktischen Restriktionen (z. B.: csak jel = nur attr) und auch mit stilistisch-pragmatischen Angaben (z. B.: kissé vál = ein wenig gewählt) versehen sind. Didaktisch relevant ist des Weiteren, dass bei verbalen Kollokatoren auch die sog. Partnerfunktion (z. B.: AKK) als Index angegeben wird und die interlingualen Kollokationen mit dem Symbol „☝“ markiert sind (siehe Abb. 1).[10]

2. Etappe II: E-KolleX DaF, das Lernerwörterbuchportal (2015–2020)

2.1 Internetwörterbuch für Kollokationen

Im ersten Band des WLWF (2010) findet man zum Terminus *Cross-Media-Publishing* weitergehende Erklärungen, die den Begriff näher spezifizieren:

> Im Anwendungsbereich der Lexikographie meint *Cross-Media-Publishing*, dass aus derselben Datenbasis verschiedene lexikographische Produkte (Printwörterbuch, CD-ROM-Wörterbuch, Online-Wörterbuch) erzeugt werden können. Während der Terminus *Cross-Media-Publishing* insbesondere auf die Spezifika einzelner Publikationsmedien abhebt, perspektiviert der bedeutungsverwandte Terminus *medienneutrales Publizieren* die Datenaufbereitung und -haltung, bei der das Datenmodell möglichst unabhängig von der Präsentation der Daten in verschiedenen Publikationsmedien bleiben sollte. (WLWF 2010, S. 751)

Zum Online-Release der Daten wurde das Redaktionssystem im Jahre 2016 von einer Desktop-Applikation auf ein web-basiertes User-Interface umgestellt und gleichzeitig in Teilen als erster Protoyp eines zweisprachigen Internetwörterbuchs für Kollokatationen für das breite Lernerpublikum kostenlos und ohne Werbung ins Netz gestellt. Die automatische Verlinkung mit bereits existierenden deutschen wissenschaftlichen Sprachressourcen hat mehrere Vorteile für ein Online-Kollokationswörterbuch: größere Zuverlässigkeit der Daten, Ökonomie im Hinblick auf sprachliche Ressourcen, gesteigerte Interaktivität durch durchdachte Vernetzung, gezielte komplementäre Effekte sowie Synergieeffekte zwischen den Ressourcen.

[10] Alle Charakteristika des Printwörterbuchs sind im deutschsprachigen Nachspann (KolleX 2014, S. 953–979) ausführlich beschrieben.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Dank der kontinuierlichen Weiterentwicklung seit 2016 ist allmählich ein Wörterbuchportal (i. S. v. Engelberg/Storrer 2016, S. 57) für Lerner entstanden, das im nächsten Abschnitt vorgestellt wird.

2.2 Deutsches Lernerwörterbuchportal

Der erste Prototyp wurde vor dem Hintergrund der Ergebnisse der Wörterbuchbenutzungsforschung für Onlinewörterbücher (vgl. Müller-Spitzer/Koplening/Töpel 2012) entworfen.[11] Im Folgenden werden die bereits (teil)realisierten Module des E-KolleX DaF in Kürze vorgestellt.

Die ersten zwei Module *KOLLEX* und *Schulwörterbuch* bilden das lexikographische Grundgerüst des Lernerwörterbuchportals, weil sie alle 1274 substantivischen Wörterbuchartikel aus dem Printwörterbuch SZÓkapTÁR/KolleX (2014) und 693 printlayout-formatierte Substantivartikel des Schulwörterbuchs Holl-Suli (2001) umfassen (siehe Tab. 2). Das zweisprachige Grundschulwörterbuch Holl-Suli wurde sukzessive retrodigitalisiert, damit lernerorientierte Beispielsätze und ihre jeweiligen Übersetzungen zu den – auch im KolleX vorhandenen – Lemmata online dargeboten werden können. Da Kollokationen in der Sprachproduktion eine zentrale Rolle spielen (vgl. Réder 2006) und in einem Lernerwörterbuchportal zum primären Wörterbuchgegenstand gehören sollten, bilden die Daten von KolleX das Herzstück des Portals. Die anhand des DW ermittelten, statistisch signifikanten und manuell sememspezifisch ausgewählten Kollokationen und lernerrelevanten Wortverbindungen mit ihren jeweiligen Übersetzungen stammen also aus der Datenbank von KolleX, wo sie nach ihren Strukturtypen geordnet und mit vielen anderen semantischen und pragmatischen Angaben gespeichert vorliegen. Sie werden in der Webseitenarchitektur unter den linear angeordneten Strukturlinks dargeboten, nebst dem ersten Tab namens *Suliszótár* (dt. *Schulwörterbuch*).

Das dritte Modul *externe Ressourcen* impliziert die Integration lemmabezogener Portale: Bei E-KolleX DaF sind diese das lexikologische Portal DW und das Wortschatzportal DWDS (vgl. Engelberg/Müller-Spitzer/Schmidt 2016). Damit wird gleichzeitig teilweise das vierte Modul *Korpus mit spezifischen Analysetools* realisiert.

Gouws (2014) stellt mit Hilfe von Screenshots von *elexiko* plastisch vor, wie dank der elektronischen Datenpräsentation eine neue Datendistribution und eine neue Artikelstruktur des Typs „dynamic multi-layered article structure“ (ebd., S. 165) entsteht:[12]

> Clicking on the structural indicator „weiter“ (= further) in anyone of these sub-blocks guides the user to a further layer in the treatment of the lemma […]. Clicking on the data-identifying entry “Grammar” guides a user to the display in Figure 5. The grammar data here would typically be part of a traditional comment on form but here it is isolated from the text block containing comment on form items […]. (ebd., S. 164)

[11] Dieser Prototyp wurde in Hollós (2019) detailliert beschrieben.

[12] Näheres zu dieser Datendistribution und Artikelstruktur im E-KolleX relativ zum Webdesign findet sich in Hollós (2018, S. 164–167).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Diese Art von Links werden bei EngelbergMüller-Spitzer/Schmidt (2016, S. 159) als Stukturund Inhaltslinks definiert. Erstere sind Elemente der inneren Zugriffsstruktur, die letzteren gehören jedoch zum Wörterbuchgegenstand.

Die neue Dynamik entsteht demgemäß auch im Portal E-KolleX DaF einerseits durch Strukturlinks und andererseits dank neuer Datentypen, realisiert durch Inhaltslinks mit verschiedenen typisierten Linkanzeigern, die auch den direkten Zugriff auf weitere Ressourcen ermöglichen.

Erstrangiges Ziel der folgenden Unterkapitel ist es dementsprechend, einige lexikographische Potenziale moderner Onlineressourcen (vgl. Lemberg 2001; Gouws 2014; Müller-Spitzer et al. 2014 u. a.), insbesondere die Dynamik am aktuellen Prototyp des E-KolleX-Portals mit seinen Modulen nachzuweisen.[13]

Diese Potenziale sind bei Lemberg (2001, S. 73) bereits früh aufgelistet und erörtert worden. Von ihnen werden hier nur drei genannt und im Folgenden exemplifiziert:

- Hypertextualisierung,
- multimediale Aufbereitung lexikographischer Daten,
- Aufhebung eines statischen zugunsten eines dynamischen Wörterbuchs.

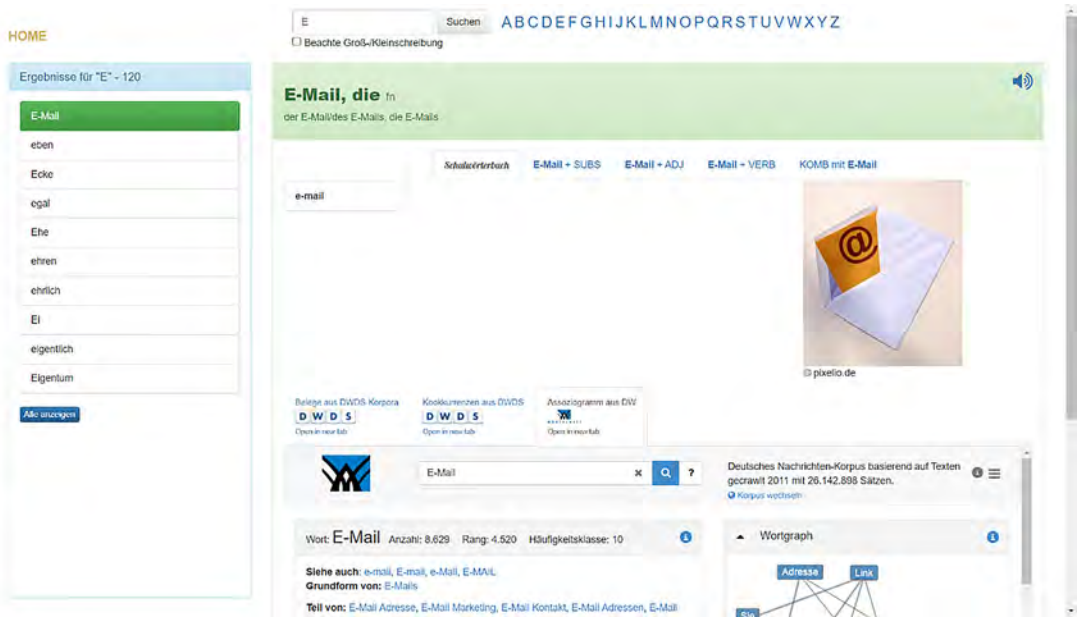
2.2.1 Dynamik durch neue Datentypen

Die Umstellung des Datenbankmanagment- und Redaktionssystems von einer Desktop-Applikation (2005–2015) zu einem webbasierten, dynamischen System (2016-) hat das erste Onlinerelease der Daten von E-KolleX ermöglicht und alle neuen Anforderungen von Lemberg (2001) bis auf die Kooperation zwischen Lexikographen und Nutzern gleichzeitig erfüllt. Neue Datentypen, wie z. B. Bild- und Tondateien, konnten in die Datenbank integriert werden, die zur Multimodalität der Ressource beigetragen haben.[14] Die neue Dynamik entstand auch durch Hypertextualisierung, indem Struktur- und Inhaltslinks im Webdesign realisiert wurden. Der nächste Screenshot zeigt diese Elemente auf der Präsentationsebene von E-KolleX DaF, indem rechts ein Bild zum Lemmazeichen und etwas rechts oberhalb ein Symbol für die Aussprache sowie oben und unten verschiedene Reiter zu sehen sind. Die unteren Tabs sind Inhaltslinks wie „Assoziogramm“ aus DW sowie „Belege“ und „Kookkurrenzen“ aus DWDS, während die oberen Strukturlinks gemäß der Strukturtypen der Kollokationen den gezielten Zugriff auf verschiedene Kollokatoren/Kollokationen ermöglichen:

[13] Da hier Kürze geboten ist, siehe ausführlich Hollós (2018, S. 156–169).

[14] Die Bereicherung des Internetwörterbuchs durch sememspezifische Bilder für die substantivischen Lemmata erfolgte systematisch am Anfang jeder Lemmastrecke. Dadurch gibt es fast 400 Bilder im Portal. Die Tondateien sind allerdings noch nicht realisiert.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.



Screenshot des Portals E-KolleX DaF mit dem Lemma **E-Mail**

Abb. 2: Neue Datentypen (Bild-, Tondatei) und verschiedene Tabs (unten: Inhaltslinks, oben: Strukturlinks) im Wörterbuchartikel zum Lemmazeichen E-Mail

Bei polysemen substantivischen Lemmazeichen wird in der ersten Lemmateilstrecke (hier: von **E-Mail** bis **Eigentum**) pro Semem jeweils ein Bild angesetzt (siehe z.B. Semem 2 zu **Nachricht** mit der Bedeutung 'kurze mündlich oder schriftliche Mitteilung' in Abb. 2). Das erleichtert einerseits die Desambiguierung des Wortes bei den Lernern und sorgt zugleich für ein noch dynamischeres multimodales Webdesign der Wörterbuchartikel.

2.2.2 Integration weiterer Ressourcen

Neue Perspektiven für das damalige Internetwörterbuch hat jedoch erst die Integration eines Schulwörterbuchs und die der bereits erwähnten einsprachigen lemmabezogenen Portale wie DW und DWDS eröffnet. Letztere sorgen für eine monodirektionale Vernetzungsstruktur mit (teilweise) direktem Zugriff auf bestimmte Datentypen in den ausgewählten Onlineressourcen (vgl. Hollós 2018). Gleichzeitig werden die Lemmazeichen mit Korpora verlinkt. Dies ermöglicht wiederum die gezielte Recherche nach bestimmten morphologischen, syntaktischen und semantischen Eigenschaften der nachgeschlagenen Wörter in Belegen.

Für die Integration des Schulwörterbuchs HOLL-SULI (2001), das für ungarischsprachige Deutschlerner zuerst als Printwörterbuch erschien, mussten alle Wörterbuchartikel zuerst aktualisiert werden, damit 2017 mit der Eingabe der Substantivartikel in die Datenbank begonnen werden konnte.

Der folgende Screenshot zeigt nicht nur den retrodigitalisierten Schulwörterbuchartikel, sondern unter Semem 2 auch einen weiteren Link „**Nachrichten**", der für diesen Artikel aus dem PONS Bilderwörterbuch manuell ausgesucht und eingefügt wurde. Dieser Schulwörterbuchartikel bedeutet gleichzeitig den Einstieg in den Portalartikel zum Lemma **Nachricht**:

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

wa_2 zu **Nachricht** im E-KolleX DaF

Abb. 3: Inhaltslinks zu weiteren Ressourcen im Wörterbuchartikel zum Lemmazeichen Nachricht

In der obigen Abbildung sind nicht nur die lernerrelevanten mikrostukturellen Charakteristika des Schulwörterbuchs ersichtlich, sondern im Postkommentar auch weitere Inhaltslinks, deren Linkanzeiger auch typographisch hervorgehoben sind und in einer eigenen Suchzone erscheinen.[15]

Das zweisprachige Internetwörterbuch (2016) und das nachfolgende, seit 2017 im Entstehen befindliche Lernerwörterbuchportal E-KolleX DaF wurden von Anfang an für die deutschlernende oder sprachinteressierte Öffentlichkeit kostenlos unter der Internetadresse www.kollex.hu/szotar in großen Teilen zugänglich gemacht.

Die entsprechenden Kennzahlen des Lernerwörterbuchportals E-KolleX DaF findet man in der nächsten Tabelle (Stand: März 2022):

Substantivlemmata	**1274**
Kollokationen/Wortverbindungen	**ca. 30000**
Wörterbuchartikel aus Holl-Suli	**693**
Bilder zu Sememen der Substantivlemmata	**269**
Dynamische Links zu Angaben in externen Ressourcen (DW, DWDS)	**3822**
Links zu weiteren externen Ressourcen (PONS-Bildwörterbuch etc.)	**ca. 2700**
Inhaltslinks insgesamt (DW, DWDS, PONS etc.)	**ca. 6500**

Zahlen von E-KolleX DaF

Tab. 2: Exakte und geschätzte Zahlen von Substantivlemmata, Wortverbindungen/Kollokationen, Links, retrodigitalisierten Schulwörterbuchartikeln, Bildern u. a. im E-KolleX DaF

[15] Diese Inhaltslinks – ähnlich wie die Bilder – werden zurzeit nur bei den Musterartikeln, d. h. bei substantivischen Lemmazeichen in der ersten Lemmateilstrecke (hier: von **Nachbar** bis **Nagel**) angesetzt.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

3. Herausforderungen der Etappe III von E-KolleX DaF

Anstatt eines Fazits wird in diesem letzten Kapitel der Versuch unternommen, einen Ausblick zu bieten, indem die möglichen Wege einer Weiterentwicklung des Portals E-KolleX DaF und die damit verbundenen Herausforderungen skizziert werden. Manche Desiderata – wie die Erweiterung der Datenbank mit Bildern und mit Inhaltslinks in den retrodigitalisierten Schulwörterbuchartikeln zu allen Substantivlemmata oder die noch fehlenden Tondateien – wurden bereits im Laufe des Artikels angedeutet.[16] Eine der wichtigsten Entscheidungen ist jedoch, wie die Weiterentwicklung erfolgen könnte, weil daraus unterschiedliche Fragestellungen resultieren: *Welche Module wurden für das Lernerwörterbuch von Anfang an geplant und bisher nicht oder nur teilweise realisiert?* oder: *Welche aktuellen Herausforderungen stellen sich für das Projekt, seitdem die Konzeption und die Datenbank entstanden sind?* Beide Fragen können an dieser Stelle nur ansatzweise beantwortet werden.

Die erste Möglichkeit der Weiterentwicklung ist dementsprechend, die nicht oder nur teilweise realisierten Module wie das Schulwörterbuch-Modul zu vervollständigen, da allein für die Substantivartikel noch ca. 600 didaktisch motivierte Schulwörterbuchartikel zu erstellen sind.[17] Wenn eine nächste Printpublikation ausbleibt, werden nicht nur die Äquivalentangaben zu den Lemmazeichen anderer Wortarten online veröffentlicht, sondern auch die Kollokatoren/Kollokationen sowie die Kombinationen dazu.

Bei der zweiten Möglichkeit wären z. B. folgende Herausforderungen zu nennen: Aktualisierung der Daten durch DWDS-Wortprofil und die korpusgesteuerte Ermittlung bzw. Integration größerer typischer Einheiten sowohl auf der Datenmodellierungs- als auch auf der Präsentationsebene.[18] Letztere bedeutet sowohl eine technische als auch lexikographische Herausforderung mit viel Arbeitsaufwand. Auch eine ein- und/oder mehrsprachige Version des Portals, d. h. ohne Zielsprache oder mit einer anderen Zielsprache als Ungarisch, könnte in Erwägung gezogen werden.

Zuletzt könnte – im Sinne des zum Titel gewählten Terminus *Cross-Media-Publishing* – über die Integration in ein größeres, ungarisch- oder deutschsprachiges Portal sowie über eine App-Version nachgedacht werden.

Welcher Weg der Datenbank des Lernerwörterbuchportals E-KolleX DaF in der Zukunft beschieden wird, werden letztendlich die Community und die Ressourcen entscheiden.

Literatur

Belica, C. (1995): Statistische Kollokationsanalyse und -clustering. Korpuslinguistische Analysemethode. Mannheim. http://corpora.ids-mannheim.de/ (Stand: 23.3.2022).

DW: Deutscher Wortschatz. https://wortschatz.uni-leipzig.de/de (Stand: 23.3.2022).

[16] Folgende erste Herausforderungen sind hier noch zu nennen: Alle Umtexte müssten aktualisiert, dem Onlineformat angepasst, schließlich mit einem entsprechenden Webdesign im Portal integriert werden. Neuentwickelte Benutzungstutorials könnten den Benutzer auf dem Weg zur erfolgreichen Konsultation unterstützen und damit die Lernerautonomie sowie den Lerneffekt erhöhen.

[17] Auch die Schulwörterbuchartikel zu den Verb-, Adjektiv- und Adverblemmata müssten im Portal retrodigitalisiert werden, wie es z. B. beim Verb *backen* oder *baden* bereits geschehen ist.

[18] Dies konnte in der Zeit der Datenselektion und -aufbereitung zwischen 2005 und 2009 mangels eines geeigneten Analysetools nur ansatzweise im Datentyp „KOMBINATION“ (siehe Abb. 1) realisiert werden. Davon gibt es zurzeit in der Datenbank 2.261 (siehe Tab. 1).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

DWDS: Digitales Wörterbuch der Deutschen Sprache. Das Wortauskunftssystem zur deutschen Sprache in Geschichte und Gegenwart. https://www.dwds.de (Stand: 23.3.2022).

E-KolleX DaF: Elektronisches Kollokationslexikon Deutsch als Fremdsprache. http://kollex.hu/szotar/ (Stand: 23.3.2022).

elexiko: https://www.owid.de/docs/elex/start.jsp (Stand: 23.3.2022).

Engelberg, S./Müller-Spitzer, C./Schmidt, T. (2016): Vernetzungs- und Zugriffsstrukturen. In: Klosa, A./Müller-Spitzer, C. (Hg.): Internetlexikografie. Ein Kompendium. Berlin/Boston., S. 153–195.

Engelberg, S./Storrer, A. (2016): Typologie von Internetwörterbüchern und -portalen. In: Klosa, A./Müller-Spitzer, C. (Hg.): Internetlexikografie. Ein Kompendium. Berlin/Boston, S. 31–63.

Geyken, A. (2011): Statistische Wortprofile zur schnellen Analyse der Syntagmatik in Textkorpora. In: Abel, A./Zanin, R. (Hg.): Korpora in Lehre und Forschung. Bozen, S. 115–137.

Gouws, R. H. (2014): Article Structures: Moving from Printed to e-Dictionaries. In: Lexikos 24, S. 155–177.

Hausmann, F. J. (1984): Wortschatzlernen ist Kollokationslernen. Zum Lehren und Lernen französischer Wortverbindungen. In: Praxis des neusprachlichen Unterrichts 31, S. 395–406.

Hollós, Z. (2004): Lernerlexikographie: syntagmatisch. Konzeption für ein deutsch-ungarisches Lernerwörterbuch. (= Lexicographica. Series Maior 116). Tübingen.

Hollós, Z. (2008): Kollokationen und weitere typische Mehrwortverbindungen in der ungarischen Lexikographie. In: Hausmann, F. J. (Hg.): Collocations in European lexicography and dictionary research. Kollokation in der europäischen Lexikographie und Wörterbuchforschung. (= Lexicographica 24). Tübingen, S. 121–133.

Hollós, Z. (2016): Korpusbasierte intra- und interlinguale Kollokationen. In: Corpas Pastor, G. (Hg.): Computerised and corpus-based approaches to phraseology: Monolingual and multilingual perspectives (full papers) – Fraseología computacional y basada en corpus: perspectivas monolingües y multilingües (Trabajos completos). Genf, S. 302–315. http://www.tradulex.com/varia/Europhras2015.pdf (Stand: 23.3.2022).

Hollós, Z. (2018): Datendistribution relativ zum Webdesign. Der erste Protoyp des E-KolleX. In: Jesenšek, V./Enčeva, M. (Hg.): Wörterbuchstrukturen zwischen Theorie und Praxis. (= Lexicographica. Series Maior 154). Berlin, S. 151–171.

Hollós, Z. (2019): Prototyp eines zweisprachigen Internetwörterbuchs für DaF. In: Lexicographica 34, S. 65–88.

HOLL-SULI (2001): Hollós, Z.: Német-magyar Suliszótár. Második, javított és bővített kiadás. 2. Auflage. Szeged.

Klosa, A. (2007): Korpusgestützte Lexikographie: besser, schneller, umfangreicher. In: Kallmeyer, W./Zifonun, G. (Hg.): Sprachkorpora – Dantemengen und Erkenntnisfortschritt. (= Jahrbuch des Instituts für Deutsche Sprache 2006). Berlin/New York, S. 105–122.

Klosa, A./Müller-Spitzer, C. (Hg.) (2016): Internetlexikografie. Ein Kompendium. Berlin/Boston.

Lemberg, I. (2001): Aspekte der Online-Lexikographie für wissenschaftliche Wörterbücher. In: Lemberg, I./Schröder, B./Storrer, A. (Hg.): Chancen und Perspektiven computergestützter Lexikographie. Tübingen, S. 71–91.

Müller-Spitzer, C./Koplenig, A./Töpel, A. (2012): Online dictionary use: key findings from an empirical research project. In: Granger, S./Paquot, M. (Hg.): Electronic lexicography. Oxford, S. 425–457.

PONS Bildwörterbuch: www.bildworterbuch.com (Stand: 23.3.2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Quasthoff, U. (2009): Korpusbasierte Wörterbucharbeit mit den Daten des Projekts *Deutscher Wortschatz*. In: Linguistik online 39, 3/09.

Réder, A. (2006): Kollokationen in der Wortschatzarbeit. Wien.

KolleX (2014): Hollós, Z.: SZÓkapTÁR. Német–magyar SZÓkapcsolatTÁR. Korpuszalapú kollokációs tanulószótár. KolleX: deutsch-ungarisches KOLLokationsLEXikon. Korpusbasiertes Wörterbuch der Kollokationen. Deutsch als Fremdsprache. Szeged.

VALBU (2004): Schumacher, H./Kubczak, J./Schmidt R./de Ruiter, V.: VALBU – Valenzwörterbuch deutscher Verben. (= Studien zur Deutschen Sprache 31). Tübingen.

WLWF (2010): Wörterbuch zur Lexikographie und Wörterbuchforschung. Dictionary of lexicography and dictionary research. Bd. 1. Hrsg. und bearbeitet v. Wiegand, H. E. et al. Berlin/New York.

Kontaktinformationen

Zita Hollós
Károli Gáspár University of the Reformed Church in Hungary (Budapest)
hollos.zita@kre.hu

Danksagung

Mein besonderer Dank gilt Uwe Quasthoff, dem Leiter des Projekts *Deutscher Wortschatz*, der die statistisch ermittelten Kookkurrenzlisten dem Projekt zwischen 2005–2009 zur Verfügung gestellt hat. Ich danke auch Cyril Belica am IDS, der den studentischen Hilfskräften bei der zweiten Datenselektion zwischen 2009–2010 den uneingeschränkten Zugriff auf die CCDB ermöglicht hat. Ich möchte auch dem *Ungarischen Nationalfonds für Wissenschaft und Forschung* „OTKA“ (2005–2008) und der *Alexander von Humboldt Stiftung* (2009, 2012, 2013, 2016) meinen Dank aussprechen, die sowohl das Forschungsvorhaben als auch die Realisierung des Projekts finanziell maßgeblich unterstützt haben.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

(Promoting) Dictionary Use

Andrea Abel

WÖRTERBÜCHER DER ZUKUNFT IN BILDUNGSKONTEXTEN DER GEGENWART

Eine Fallstudie aus dem Südtiroler Schulwesen

Abstract The focus of this paper will be on lexical information systems and the framework guidelines for the definition of the curricula within the educational system of the Autonomous Province of Bolzano/Bozen (Italy). In Italy, the competences to be achieved at different school levels are published in the form of general guidelines. On this basis each school has to specify the general competency goals and to spell them out in a concrete curriculum.

In this paper I will examine to what extent lexical information systems are represented in the framework guidelines within the German and the Italian educational system of the Autonomous Province, these being separate systems. In a second step, I will check the representations of the resources against the "Villa Vigoni Theses on Lexicography". Finally, I will discuss the results and give an outlook for further research.

Keywords Lexikalisches Informationssystem; Wörterbuchbenutzung; Wörterbücher in der Schule

1. Einleitung

Die Villa-Vigoni-Thesen 2018 definieren „Wörterbücher der Zukunft" als „lexikalische bzw. sprachliche Informationssysteme, in denen die bestehenden lexikografischen Daten zusammengeführt sind, in denen Mehrsprachigkeit und sprachliche Varietät verankert sind und in denen die Menschen bei Wissenslücken eine Antwort sowie Unterstützung in Schreib- und Formulierungsprozessen von Texten finden" (Villa-Vigoni-Thesen 2018). Einige der 15 Thesen nehmen Bezug auf die Relevanz lexikalischer Informationssysteme in Bildungskontexten. Dieser Beitrag widmet sich anhand einer Fallstudie der Rolle solcher Informationssysteme bzw. von Wörterbüchern im schulischen Umfeld der Autonomen Provinz Bozen – Südtirol (Italien). Dabei stehen zwei Fragen im Vordergrund: Inwiefern sind lexikalische Informationssysteme in den Rahmenrichtlinien der Schule mit deutscher sowie italienischer Unterrichtssprache repräsentiert? Inwiefern spiegeln die Darstellungen Inhalte der Villa-Vigoni-Thesen 2018 wider?

2. Das Südtiroler Schulwesen

Die Fallstudie bezieht sich auf das Schulsystem der offiziell mehrsprachigen Autonomen Provinz Bozen – Südtirol. Deren Schulsystem folgt im Allgemeinen den nationalen Regelungen, weist aber auf der Grundlage des Zweiten Autonomiestatuts von 1972 (Das neue Autonomiestatut 2019/1972) eine Reihe von Besonderheiten auf. So gibt es etwa drei verschiedene Schulämter, ein deutsches, ein italienisches und ein ladinisches, die getrennt voneinander organisiert sind. Außerdem besteht das Recht auf erstsprachlichen Unterricht in den drei genannten Sprachen, die in der Provinz den Status von Amtssprachen besitzen. Daneben besteht die Pflicht, Deutsch resp. Italienisch als zweite Sprache zu unterrichten.[1] Englisch

[1] Eine Ausnahme bildet das Ladinische, die Sprache der kleinsten Minderheit in Südtirol. Im Folgenden wird das Ladinische nicht weiter berücksichtigt.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

wird wie im gesamten Staatsgebiet auf allen Schulstufen als Fremdsprache unterrichtet. Den nationalen Regelungen entspricht auch die Unterteilung des Schulsystems in eine fünfjährige Primarstufe und eine dreijährige Sekundarstufe I, die einheitlich für alle Schüler*innen ist, sowie eine fünfjährige Sekundarstufe II, deren Abschluss den Hochschulzugang ermöglicht. Die Schulen in Italien besitzen ein hohes Maß an Autonomie. Staatlich vorgegeben werden Rahmenrichtlinien, die die Kompetenzziele für jede Schulstufe beschreiben, wobei für Südtirol einige sprachbezogene Besonderheiten gelten. Die an die lokale Situation angepassten Rahmenrichtlinien für die sprachlich getrennten Schulsysteme werden entsprechend getrennt voneinander entwickelt.[2] Jede Schule ist für die Erstellung darauf basierender, konkreter Curricula selbst zuständig.

3. Datengrundlage und Methode

Für die Studie wurden im Rahmen einer Dokumentenanalyse (vgl. Mayring 2016, S. 46–50) die Südtiroler Rahmenrichtlinien der deutschen und der italienischen Schule aller drei Bildungsstufen für die Sprachen Deutsch und Italienisch als Erstsprache (L1) sowie als Zweitsprache (L2) und für Englisch als Fremdsprache (L3/EN) untersucht (Rahmenrichtlinien DE, Rahmenrichtlinien IT). Die Rahmenrichtlinien geben die allgemeinen Kompetenzziele, die in Form von Deskriptoren für Fertigkeiten und Kenntnisse dargestellt sind, nicht nur für jede der drei Schulstufen, sondern auch für die jeweiligen Zwischenstufen vor. Demnach ergibt sich folgende Untergliederung der Stufen und den entsprechenden Kompetenzzielen:

- Primarstufe: Triennium (1., 2. und 3. Klasse) und Biennium (4. und 5. Klasse) im deutschen Schulsystem vs. Biennium (1. und 2. Klasse) und Triennium (3., 4. und 5. Klasse) im italienischen Schulsystem
- Sekundarstufe I: Biennium (1. und 2. Klasse) und Monoennium (3. Klasse) in beiden Schulsystemen
- Sekundarstufe II: 1. Biennium (1. und 2. Klasse), 2. Biennium (3. und 4. Klasse) und 5. Klasse in beiden Schulsystemen[3]

Die Sekundarstufe II umfasst die Gymnasien sowie die Fachoberschulen mit den jeweils spezifischen Fachrichtungen und Fächern. Dabei sind die Richtlinien für die L1, die L2 und die L3 identisch für alle Schultypen und Fachrichtungen identisch. Erwähnenswert ist außerdem, dass die Rahmenrichtlinien des deutschen und des italienischen Bildungssystems getrennt voneinander entwickelt und formuliert werden, es sich folglich nicht um Übersetzungen handelt. Sie stehen für beide Schulsysteme in der jeweiligen Unterrichtssprache zur Verfügung.

Im Zuge der Dokumentenanalyse wurden die Deskriptoren der einzelnen Stufen und Zwischenstufen für die beiden Schulsysteme und die drei Sprachfächer L1, L2 und L3 analysiert und inhaltsanalytisch qualitativ ausgewertet (vgl. Mayring 2016, S. 114–121). Dabei wurden all jene Deskriptoren berücksichtigt, die Hinweise auf Wörterbücher oder andere sprachliche Ressourcen im weitesten Sinne enthalten, ausgehend von der Definition von „Wörterbüchern der Zukunft“ in den Villa-Vigoni-Thesen 2018, die im Folgenden verschiedenste

[2] Die Rahmenrichtlinien stehen in den Unterrichtssprachen der Schulen öffentlich online zur Verfügung; für die italienischen Rahmenrichtlinien ist auch eine deutsche Übersetzung online vorhanden, die in diesem Beitrag nicht berücksichtigt wird.

[3] Die Bezeichnungen der Zwischenstufen sind den Rahmenrichtlinien entnommen.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Formen lexikalischer Ressourcen umfassen sollen – von einfachen Wörterbüchern bis hin zu umfassenden Informationssystemen. Eine zusammenfassende Darstellung der Ergebnisse wurde an Mitarbeiter*innen der Schulbehörden mit Zuständigkeiten in sprachlichen Bereichen mit der Bitte um kritische Rückmeldung zugesandt, um einen Abgleich mit deren Einschätzungen vornehmen zu können.[4]

In einem zweiten Schritt fand ein Abgleich mit den Villa-Vigoni-Thesen statt, wobei die Bezüge zu Bildungskontexten berücksichtigt werden. Die Bezugnahme auf die Villa-Vigoni-Thesen erschien aus mehreren Gründen sinnvoll: Zum einen gehen sie – auch mit Blick auf die Zukunft – von einer breiten Auffassung von Wörterbüchern aus, die eine Entwicklung weg vom traditionellen Wörterbuchbegriff aufgreift (vgl. Levy/Steel 2015, S. 178). Zum anderen enthalten sie explizite Hinweise auf Bildungskontexte.

4. Ergebnisse

4.1 Lexikalische Informationssysteme in den Rahmenrichtlinien

4.1.1 Terminologie

Aus der Dokumentenanalyse geht hervor, dass lexikografische Ressourcen eine wichtige Rolle in den Rahmenrichtlinien sowohl der Schule mit deutscher als auch der Schule mit italienischer Unterrichtssprache spielen. Sie werden in allen Schulstufen beider Schulsysteme berücksichtigt. Dabei lässt sich feststellen, dass einerseits explizit auf lexikalische Informationssysteme hingewiesen wird, andererseits implizit. Als implizite Hinweise sollen hier solche gelten, bei denen nicht eindeutig ist, inwiefern sie sich auf lexikalische Informationssysteme beziehen. In beiden Fällen, also auf Deutsch und Italienisch, werden eine ganze Reihe unterschiedlicher Bezeichnungen verwendet.

Als explizite Hinweise lassen sich folgende Bezeichnungen finden:
- „Nachschlagewerk(e)“ / „strumenti di consultazione“
- „Mittel zum Nachschlagen“ / „mezzi di consultazione“
- „Wörterbuch“ / „dizionario“
- „Lehrmittel zum Nachschlagen“
- „Nachschlagewerke – auch digitale“
- „Lexika“
- „Dizionari ed enciclopedie su supporto sia cartaceo sia digitale“ [sowohl gedruckte als auch digitale Wörterbücher und Enzyklopädien[5]]
- „Enciclopedia multimediala“ [multimediale Enzyklopädie]
- „Strumenti di consultazione, compresi quelli multimediali“ [Nachschlagewerke, auch multimediale]
- „Dizionari monolingue e bilingue, compresi quelli multimediali“ [einsprachige und zweisprachige Wörterbücher, auch multimediale]

[4] Rückmeldungen vonseiten der Pädagogischen Abteilung und des Schulinspektorats der Deutschen Bildungsdirektion sowie der Landesdirektion italienischsprachige Grund-, Mittel- und Oberschulen staatlicher Art der Italienischen Bildungsdirektion

[5] Eigene Übersetzungen werden in eckigen Klammern angegeben. In den Fällen, in denen die in den Rahmenrichtlinien vorgefundenen deutschen und italienischen Bezeichnungen Übersetzungsäquivalente darstellen, werden sie lediglich durch Schrägstrich voneinander getrennt, ohne eine zusätzliche Übersetzung hinzuzufügen.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

- „Dizionari cartacei e *on line*“[6] [gedruckte und Online-Wörterbücher]
- „Vocabulari“ [Wörterbücher]
- „Dizionari di diverso tipo (monolingue, bilingue, *on line*)“ [verschiedene Arten von Wörterbüchern (einsprachig, zweisprachig, online)]
- „Dizionari e risorse *on line*“ [Wörterbücher und Online-Ressourcen]
- „Glossari relativi ad argomenti settoriali“ [Fachglossare]
- „Strumenti di consultazione cartacei (enciclopedie, dizionari, glossari relativi ad argomenti settoriali ecc.)“ [gedruckte Nachschlagewerke (Enzyklopädien, Wörterbücher, Fachglossare etc.)]

Implizite Hinweise hingegen sehen folgendermaßen aus:
- „Informationsquellen“
- „Internet“
- „Quellen“
- „Hilfsmittel“
- „Sprachmittel“
- „Fachspezifische Werke“
- „Englische Webseiten“
- „fonti digitali e cartacee“ [digitale und gedruckte Quellen]
- „supporti anche digitali“ [Unterstützung/Hilfsmittel, auch digital]
- „materiali di supporto anche digitale“ [Hilfsmittel, auch digital]
- „sussidi di vario tipo“ [unterschiedliche Hilfsmittel]

(Rahmenrichtlinien DE, Rahmenrichtlinien IT)

Bei den Hinweisen fällt auf, dass die deutschen Dokumente weniger unterschiedliche Begriffe verwenden und zu allgemeineren Bezeichnungen tendieren, während die italienischen Dokumente mehr unterschiedliche und eher präzisere Bezeichnungen (z. B. durch die Verwendung von präzisierenden Attributen) bevorzugen. So gibt es etwa in den deutschen Rahmenrichtlinien keinerlei Hinweise darauf, ob ein- und/oder zweisprachige bzw. gedruckte und/oder Online-Ressourcen[7] eingesetzt werden sollen. Ausschließlich in den italienischen Rahmenrichtlinien werden zudem Enzyklopädien genannt und spezifizierende Adjektive wie „multimedial“ oder „online“ verwendet. Die deutschen Dokumente verwenden im Hinblick auf die Sekundarstufe II als spezifische Bezeichnung nur „Nachschlagewerk“, nicht hingegen „Wörterbuch“ oder „Lexikon“, die nur in den Rahmenrichtlinien der unteren Stufen aufscheinen.

Insgesamt bietet sich ein Bild einer relativ zufälligen Verwendung von Termini zur Bezeichnung unterschiedlicher lexikalischer Informationssysteme. So werden Begriffe wie „Wörterbuch“, „Nachschlagewerk“, „Lexikon“ oder „Lehrmittel zum Nachschlagen“ verwendet, ohne dass klar wird, inwiefern von Synonymen auszugehen ist (vgl. Rahmenrichtlinien DE, Rahmenrichtlinien IT).

4.1.2 Verwendungskontexte

Nach der Erfassung der verwendeten Terminologie zur Bezeichnung lexikalischer Informationssysteme wurden diese in Bezug auf die verwendeten Kontexte (Schulsystem, Schulstu-

[6] Kursivschreibung wie im Original.

[7] Ein einziges Mal wird explizit auch „digital“ als Möglichkeit erwähnt.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

fen, Sprachfächer) analysiert. Die folgenden Auszüge aus den Deskriptoren (Fertigkeiten sowie Kenntnisse) sollen der Veranschaulichung dienen:

Rahmenrichtlinien der Schulen mit deutscher Unterrichtssprache:
- „Wörterbuch“ (GS L1 40[8])
- „Vorbereitete Lehrmittel zum Nachschlagen“ (GS L2 47, 48)
- „Vorbereitete Lehrmittel zum Nachschlagen“ (GS L2 47, 48)
- „Nachschlagewerke“ (MS L2 49)
- „Nachschlagewerke und fachspezifische Werke (MS L2 51)
- „Nachschlagewerke verwenden“, „Wörterbücher, Lexika“ (MS EN 52)
- „dem Internet und anderen Quellen Informationen entnehmen“, „Nachschlagewerke, Englische Webseiten“ (MS EN 56)
- „geeignete Sprachmittel“ (OS EN 51)
- „Nachschlagewerken – auch digitalen – Informationen über Bedeutung, Aussprache, Grammatik und Rechtschreibregeln entnehmen“, „Aufbau, Zeichenerklärung und Lautschrift von Nachschlagewerken“ (OS EN 52)

Rahmenrichtlinien der Schulen mit italienischer Unterrichtssprache:
- „Utilizzare diverse strategie e strumenti per fare ipotesi su parole non note e comprenderne il significato (a partire dal contesto, osservando la somiglianza tra le parole, utilizzando il dizionario) [verschiedene Strategien und Hilfsmittel verwenden, um Hypothesen zu bilden über unbekannte Wörter und deren Bedeutung zu verstehen (indem man, ausgehend vom Kontext, Ähnlichkeiten zwischen Wörtern beobachtet und das Wörterbuch benutzt)]“, „principali tipi di informazione contenuti nel dizionario, simboli e abbreviazioni [grundlegende Arten von Informationen, die im Wörterbuch vorhanden sind, Symbole und Abkürzungen]“ (GS L1 78)
- „Usare il dizionario tipo per individuare le principali informazioni presentate sulle singole voci e per scoprire il significato e l’etimologia delle parole“ [ein Standardwörterbuch verwenden, um grundlegende Informationen in den einzelnen Einträgen zu erkennen und um die Bedeutung und die Etymologie der Wörter zu erfassen], „Principali tipi di informazione contenuti nel dizionario: alcuni simboli e abbreviazioni [grundlegende Arten von Informatuonen in einem Wörterbuch: einige Symbole und Abkürzungen]“ (GS L1 78)
- „ricavare informazioni da un dizionario o da un’enciclopedia multimediale [einem Wörterbuch oder einer multimedialen Enzyklopädie Informationen entnehmen]“ (GS L2 85)
- „Usare dizionari di vario tipo per individuare le diverse informazioni presentate sulle singole voci, per l’autocorrezione, per risolvere dubbi linguistici e per scoprire l’etimologia delle parole“ [Wörterbücher unterschiedlicher Art verwenden, um verschiedene Informationen in einzelnen Einträgen zu erkennen, zur Selbstkorrektur, um sprachliche Zweifel zu lösen und die Etymologie von Wörtern zu entdecken], „Le informazioni contenute nel dizionario: simboli e abbreviazioni“ [Informationen im Wörterbuch: Symbole und Abkürzungen] (MS L1 80)
- „Utilizzare in modo efficace i dizionari monolingue e bilingue, compresi quelli multimediali“ [die grundlegenden ein- und zweisprachigen Wörterbücher effizient verwenden, einschließlich multimedialer], „Dizionari cartacei e *on line* e loro tecniche d’uso“ [gedruckte und Online-Wörterbücher und deren Verwendungsweisen] (OS L1 20)

[8] Legende: GS = Grundschule (Primarstufe), MS = Mittelschule (Sekundarstufe I), OS = Oberschule (Sekundarstufe II); Sprachfächer L1, L2, L3/EN; Seitenzahl Rahmenrichtlinien.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

- „Essere in grado di comprendere, analizzare e, se necessario, interpretare testi autentici diversi – anche riportati dai media – o appartenenti a generi letterari diversi, opere letterarie complete o testi estratti da opere letterarie, senza o con l'aiuto dei dizionari" [imstande sein, mit oder ohne die Hilfe von Wörterbüchern unterschiedliche authentische Texte – auch aus den Medien – zu verstehen, zu analysieren und, falls nötig, zu interpretieren, auch Texte unterschiedlicher literarischer Genres, gesamte literarische Werke oder Auszüge daraus], „dizionari di diverso tipo (monolingue, bilingue, on line ...)" [Wörterbücher unterschiedlicher Art (einsprachig, zweisprachig, online ...)] (OS L2 24)
- „Utilizzare vari strumenti di consultazione e di ricerca in modo appropriato, comprese le nuove tecnologie dell'informazione e della comunicazione" [unterschiedliche Nachschlage- und Suchwerkzeuge angemessen verwenden, einschließlich der neuen Informations- und Kommunikationstechniken], „Strumenti di consultazione cartacei (enciclopedie, dizionari, glossari relativi ad argomenti settoriali ecc.) e loro tecniche d'uso" [gedruckte Nachschlagewerke (Enzyklopädien, Wörterbücher, Fachglossare etc.)] (OS EN 30)

(Rahmenrichtlinien DE, Rahmenrichtlinien IT).

Die Auszüge verdeutlichen unterschiedliche Herangehensweisen in der Art der Formulierung der Deskriptoren: Die deutschen Rahmenrichtlinien sind viel knapper gehalten als die italienischen, die viel mehr ins Detail gehen. Zudem bestehen deutliche Unterschiede zwischen den beiden Schulsystemen, sowohl was die Schulstufen als auch was die verschiedenen Sprachfächer betrifft: So werden lexikografische Ressourcen z. B. im Hinblick auf die deutschen Schulen öfter in den unteren Schulstufen, für die italienischen hingegen öfter in den höheren Schulstufen explizit erwähnt. Während Wörterbücher in den Rahmenrichtlinien der italienischen Schulen insbesondere im Zusammenhang mit der L1, gefolgt von der L2 genannt werden, spielen sie in denen der deutschen Schulen kaum eine Rolle im Kontext des L1-Unterrichts und werden dort hingegen in Bezug auf die L3 am häufigsten erwähnt (siehe Abb. 1).

Aus den Deskriptoren lässt sich zwischen den Schulstufen in beiden Schulsystemen teilweise eine leichte Progression ablesen, wie die folgenden Beispiele aus den deutschen Rahmenrichtlinien verdeutlichen: „Vorbereitete Lehrmittel zum Nachschlagen" (GS L2 47, 48) → „Nachschlagewerke" (MS Klasse 1+2 L2 49) → „Nachschlagewerke und fachspezifische Werke (MS Klasse 3 L2 51). Während in der Primarstufe I im Hinblick auf die L2 der Fokus auf vorbereiteten Lehrmitteln liegt, werden für die ersten beiden Klassen der Sekundarstufe I allgemein Nachschlagewerke ohne jegliche Anpassung genannt und schließlich für die dritte und letzte Klasse der Sekundarstufe I ein Fachbezug hinzugefügt. Eine Progression ist in den deutschen Rahmenrichtlinien sowohl für die L2 als auch für die L3, nicht hingegen die L1 erkennbar, die insgesamt eine vergleichsweise marginale Rolle einnimmt, was Hinweise auf lexikalische Ressourcen betrifft. Die folgenden Auszüge aus den italienischen Rahmenrichtlinien illustrieren eine Progression in den Deskriptoren für die L1, ähnliche Beispiele lassen sich aber auch für die L2 und die L3 finden: „Utilizzare diverse strategie e strumenti per fare ipotesi su parole non note e comprenderne il significato (a partire dal contesto, osservando la somiglianza tra le parole, utilizzando il dizionario) [verschiedene Strategien und Hilfsmittel verwenden, um Hypothesen zu bilden über unbekannte Wörter und deren Bedeutung zu verstehen (indem man, ausgehend vom Kontext, Ähnlichkeiten zwischen Wörtern beobachtet und das Wörterbuch benutzt)]" (GS L1 78) → „Usare dizionari di vario tipo per individuare le diverse informazioni presentate sulle singole voci, per l'autocorrezione, per risolvere dubbi linguistici e per scoprire l'etimologia delle parole"

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

[Wörterbücher unterschiedlicher Art verwenden, um verschiedene Informationen in einzelnen Einträgen zu erkennen, zur Selbstkorrektur, um sprachliche Zweifel zu lösen und die Etymologie von Wörtern zu entdecken] (MS L1 80) → „Utilizzare in modo efficace i dizionari monolingue e bilingue, compresi quelli multimediali“ [die grundlegenden ein- und zweisprachigen Wörterbücher effizient verwenden, einschließlich multimedialer] (OS L1 20). In der Primarstufe I wird von einer ersten Annäherung an Wörterbücher ausgegangen, wobei verschiedene Strategien angewandt werden sollen, um Rückschlüsse auf Wortbedeutungen zu ziehen. Auf der nächsten Schulstufe, der Sekundarstufe I, dienen Wörterbücher bereits dazu, konkrete sprachliche Probleme oder Zweifel zu lösen. Für die Sekundarstufe II schließlich ist eine effiziente Benutzung unterschiedlicher Wörterbucharten vorgesehen. Ein auffallendes Detail stellen die expliziten Hinweise auf die Wortetymologie im Zusammenhang mit der Wörterbucharbeit in den italienischen Rahmenrichtlinien für die L1 sowohl in der Primarstufe als auch in der Sekundarstufe I dar, die in den deutschen Rahmenrichtlinien hingegen an keiner Stelle erwähnt werden.

			Ital. Schule	Dt. Schule
Grundschule				
	L1	1+2(+3)	/	X
		(3+)4+5	X X	Y
	L2	1+2(+3)	/	X
		(3+)4+5	X	X
	EN	1+2(+3)	/	/
		(3+)4+5	/	/
Mittelschule				
	L1	1+2	Y X	/
		3	X	/
	L2	1+2	/	Y
		3	X	Y
	EN	1+2	X	X
		3	/	Y
Oberschule				
	L1	1+2	Y X	/
		3+4	X	/
		5	X	/
	L2	1+2	X	/
		3+4	X X	/
		5	X X	/
	EN	1+2	X	Y X
		3+4	X	/
		5	X	/

Abb. 1: Hinweise auf lexikalische Ressourcen im deutschen und im italienischen Schulsystem in Südtirol: Anzahl von expliziten Hinweisen (X) und impliziten Hinweisen (Y) bzw. Hinweise auf Null-Vorkommen (/) auf den einzelnen Schulstufen bzw. Zwischenstufen (die Zahlen 1–5 geben die Klassen der einzelnen Zwischenstufen wieder; in Klammern stehen diejenigen, die in den beiden Schulsystemen unterschiedlichen Zwischenstufen zugeordnet sind, siehe dazu Abschn. 3) und den einzelnen Sprachfächern (L1, L2, L3/EN)

4.3 Bezüge zu den Villa-Vigoni-Thesen 2018

Zwischen den Rahmenrichtlinien und den Villa-Vigoni-Thesen 2018 lassen sich einige, wenn auch recht marginale, Bezüge herstellen. These 1 weist keinen expliziten Hinweis zu Bildungskontexten auf, kann aber dennoch mit ihnen in Bezug gesetzt werden, geht es doch um Antworten auf „Wissenslücken“ und „Unterstützung in Schreib- und Formulierungs-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

prozessen von Texten", die lexikalische Informationssysteme geben sollen. These 14 und 15 hingegen sprechen direkt den Bereich Bildung an, einerseits indem die Nutzung digitaler sprachlicher Ressourcen als „strategische Schlüsselkompetenz" benannt wird und eine Verankerung dieser Kompetenz auch in der „Aus- und Weiterbildung von LehrerInnen" gefordert wird (These 14), andererseits indem „pädagogische Konzepte [zur] Didaktisierung lexikografischer Informationssysteme" postuliert werden, die in Zusammenhang mit einer allgemeinen „Medienkompetenz der BenutzerInnen" gestellt werden (These 15).[9]

Bei einem Abgleich mit den Villa-Vigoni-Thesen 2018 fällt auf, dass Wörterbuchressourcen in den Rahmenrichtlinien beider Schulsysteme vorwiegend mit sprachlicher Rezeption und weniger mit Produktion in Verbindung gebracht werden (vgl. These 1). Nicht klar geht aus den Rahmenrichtlinien hervor, inwiefern Hinweise auf Aspekte von Medienkompetenz – die durchaus vorhanden sind, auch als allgemeine Schlüsselkompetenz und unabhängig von einzelnen Fächern – und damit verbunden auf den Umgang mit (digitalen) Datenbanken auch lexikalische Informationssysteme inkludieren. Dasselbe gilt für Hinweise auf allgemeine Suchfertigkeiten (vgl. Thesen 14 und 15). Diesbezüglich ist die in den Rahmenrichtlinien verwendete Terminologie insgesamt relativ vage gehalten, die wiederum mehr Spielraum für die Ausformulierung der Curricula an den einzelnen Schulen lässt.

5. Diskussion und Ausblick

Die Analysen haben gezeigt, dass lexikalische Informationssysteme im Südtiroler Bildungssystem, soweit dies aus den Rahmenrichtlinien sowohl der Schule mit deutscher als auch der Schule mit italienischer Unterrichtssprache hervorgeht, die sich freilich immer an den nationalen Vorgaben orientieren, einen hohen Stellenwert einnehmen. Sie werden für alle Schulstufen und die Sprachfächer L1, L2 und L3 (EN) berücksichtigt. Auch geben die Deskriptoren in weiten Teilen eine Progression über die Schulstufen hinweg zu erkennen. Inwiefern Rahmenrichtlinien bzw. Lehrpläne in anderen Bildungskontexten eine solche ebenfalls widerspiegeln, lässt sich aus der äußerst eingeschränkten Forschungsliteratur dazu kaum erschließen. Eine Studie zur Analyse der schulischen Curricula in Südafrika etwa weist darauf, dass dort in den Sprachfächern ebenfalls eine Progression im Hinblick auf die Wörterbuchverwendung vorgesehen ist (vgl. Nkomo 2015, S. 99). Vergleichbar mit den Südtiroler Rahmenrichtlinien ist dabei, dass das Auffinden von Informationen zur Bedeutung und Schreibung zu den elementaren Wörterbuchbenutzungsfertigkeiten gezählt wird, während das Nutzen weiterer Angabeklassen als größere Herausforderung gewertet wird. Die Rahmenrichtlinien der Schulen mit deutscher Unterrichtssprache knüpfen die Wörterbuchbenutzungsfertigkeiten zudem teilweise an die Art der Ressource von angepassten Nachschlageressourcen über nicht angepasste bis hin zu solchen mit Fachbezug

[9] Die Thesen im Wortlaut (Hervorhebungen wie im Original): (1) „**Wörterbücher der Zukunft** sind lexikalische bzw. sprachliche Informationssysteme, in denen die bestehenden lexikografischen Daten zusammengeführt sind, in denen Mehrsprachigkeit und sprachliche Varietät verankert sind und in denen die Menschen bei Wissenslücken eine Antwort sowie Unterstützung in Schreib- und Formulierungsprozessen von Texten finden." (14) „Das digitale Datenangebot in den Informationssystemen der Zukunft muss als wichtiges Hilfsmittel des *lifelong learning* angesehen werden, so dass die **kritische Benutzung der Ressourcen** als **strategische Schlüsselkompetenz** etabliert wird. Dies muss auch in der **Aus- und Fortbildung von LehrerInnen** verankert werden." (15) „Die Lexikografie braucht **pädagogische Konzepte**, um die **Didaktisierung** lexikografischer Informationssysteme leisten zu können. Dabei soll eine Einbeziehung der Medienkompetenz der BenutzerInnen erfolgen."

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

(vgl. Abschn. 4.1.2). Dabei fällt auf, dass z. B. die Kategorie „Lernerwörterbuch" an keiner Stelle erwähnt wird, auch nicht in den Rahmenrichtlinien der Schulen mit italienischer Unterrichtssprache.

Gemeinsam ist den Südtiroler Rahmenrichtlinien eine scheinbar recht zufällig gewählte Terminologie zur Bezeichnung unterschiedlicher lexikalischer Informationssysteme. So lässt sich nicht erschließen, inwiefern einzelne Begriffe ganz bewusst gewählt wurden und inwiefern sie Synonyme darstellen oder nicht. Teilweise wird explizit zwischen gedruckten und digitalen Wörterbüchern unterschieden, teilweise nicht. Expliziter sind in diesem Zusammenhang die italienischen Rahmenrichtlinien, die sich insgesamt durch eine detailliertere Bezeichnungsstrategie (z. B. durch die Verwendung präzisierender Attribute) auszeichnen als die deutschen. Eine größere Ausführlichkeit weisen die italienischen Rahmenrichtlinien zudem in der Ausformulierung der Deskriptoren insgesamt auf. Mit der Ausführlichkeit verbunden ist u. a. auch eine explizite Bezugnahme auf die Sprachgeschichte der L1 durch Hinweise auf die Wortetymologie für die Primarstufe und die Sekundarstufe I.

Daneben lassen sich eine ganze Reihe von Unterschieden in Bezug auf die Schulstufen und die Sprachfächer (L1, L2, L3) zwischen den beiden Schulsystemen in Südtirol feststellen. Auffallend sind die relativ sparsamen Hinweise auf Wörterbücher in den deutschen Rahmenrichtlinien im Hinblick auf die L1. In den italienischen Rahmenrichtlinien hingegen ist das genau umgekehrt, d. h. die L1 liegt bei der Berücksichtigung lexikalischer Ressourcen vorn auf. Auch die Gewichtung der Ressourcen in Bezug auf die Schulstufen ist in den beiden Schulsystemen unterschiedlich: Die deutschen Rahmenrichtlinien nennen sprachliche Informationssysteme häufiger in den unteren, die italienischen hingegen häufiger in den höheren Schulstufen.

Die Darstellungen in den Rahmenrichtlinien spiegeln sich an einigen Stellen in den Villa-Vigoni-Thesen 2018 wider, wobei nur drei der insgesamt 15 Thesen im Zusammenhang mit Bildungskontexten relevant sind (Thesen 1, 14 und 15). Erwähnenswert ist die Tatsache, dass die Rahmenrichtlinien – im Unterschied zu den Thesen (vgl. These 1) – sprachliche Rezeptionsaktivitäten in den Vordergrund rücken und die Produktion eher vernachlässigen. Zudem geht aus den Rahmenrichtlinien nicht klar hervor, inwiefern der Umgang mit lexikalischen Informationssystemen als Teil einer Medienkompetenz aufgefasst wird und auch „die kritische Benutzung der Ressourcen" einbezieht (vgl. These 14 und 15). Dass diese Zusammenhänge relevant sind, zeigt u. a. Nied-Curcio (2015, S. 304 f.) mit Bezug auf die Bildungspläne für moderne Fremdsprachen an den Gymnasien in Baden-Württemberg. Ein eindeutiger Einbezug des Umgangs mit gedruckten und digitalen lexikalischen Informationssystemen in die Beschreibungen einer zu erwerbenden allgemeinen Medienkompetenz bzw. Methoden- und Recherchierkompetenz, einschließlich einer kritischen Einschätzung von Quellen und Informationen, in die Rahmenrichtlinien der Südtiroler Bildungswelt – aber auch darüber hinaus – könnte die Wörterbucharbeit, insbesondere wenn sie auch Angebote miteinbezieht, die nicht von Verlagen, Sprachakademien oder anderen Forschungseinrichtungen erstellt werden, verstärkt in den Vordergrund rücken.

Bei allen Gemeinsamkeiten zwischen den beiden Südtiroler Schulsystemen, die im Wesentlichen den nationalen Vorgaben folgen, spiegeln sich in den Rahmenrichtlinien, zumindest in Bezug auf den in diesem Beitrag behandelten Schwerpunkt, doch auch unterschiedliche schulweltliche Kulturen wider: Die Unterschiede schlagen sich beispielsweise in einer größeren Ausführlichkeit in den Beschreibungen, einschließlich der Bezugnahme auf die eige-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

ne sprachlich-kulturelle Vergangenheit (siehe Hinweise auf die Wortetymologie) in den italienischen Rahmenrichtlinien wider, die eine stärkere zentrale Regelung der schulischen Vorgaben nahelegen. Die weitaus knapper gehaltenen deutschen Rahmenrichtlinien hingegen eröffnen mehr Spielräume für die weitere autonome Ausgestaltung vonseiten der Schulen.

Die tatsächliche Praxis im Umgang mit Wörterbüchern in den Schulen lässt sich aus den Rahmenrichtlinien freilich nicht ablesen. Einen vorläufigen Abgleich in Bezug auf die schulische Praxis erlaubten Einschätzungen vonseiten der Schulbehörden. Allgemein, so die Auskünfte, werden in den Sprachfächern Wörterbücher aktiv eingesetzt, sowohl Print- als auch Online-Wörterbücher. Die Schulen handhaben die Verwendung und Anschaffung von Wörterbüchern im Rahmen ihrer Kompetenzen autonom. Generell stehen an den Schulen Sätze von Wörterbüchern in der eigenen Bibliothek zur Verfügung. Zudem liegen üblicherweise einsprachige Wörterbücher für die L1, die L2 und die L3 (EN) in den Klassen aller Schulstufen (Grund-, Mittel- und Oberschule) auf. Die Schulen geben den Eltern Kaufempfehlungen für eine Reihe von Wörterbüchern, so etwa, um nur ein Beispiel zu nennen, für das Kinderwörterbuch „Findefix“ (Fackelmann 2000), mit dem in der deutschen Schulwelt im L1-Unterricht an der Primarstufe häufig gearbeitet wird. Inwiefern die in den deutschen und italienischen Rahmenrichtlinien festgestellten Unterschiede in Bezug auf die Verwendung von Wörterbüchern auf den verschiedenen Schulstufen bzw. in den verschiedenen Sprachfächern die schulische Praxis reflektieren, konnte auf der bloßen Grundlage der externen Einschätzungen nicht festgestellt werden.

Auf diese und weitere noch offene Punkte zur Wörterbucharbeit in der Schule wird die Auswertung der Ergebnisse einer Fragebogenerhebung mit Lehrkräften der Sprachfächer L1, L2 und L3/EN Aufschluss geben. Dabei wird die tatsächliche Nutzung lexikographischer Ressourcen ermittelt, wobei mehrere Dimensionen berücksichtigt werden, u. a. welche digitalen und gedruckten Ressourcen (nicht) verwendet und aus welchen Gründen bestimmte Ressourcen (nicht) gewählt werden.

Literatur

[Rahmenrichtilinien DE] Rahmenrichtlinien der Schulen mit deutscher Unterrichtssprache: https://www.provinz.bz.it/bildung-sprache/didaktik-beratung/rahmenrichtlinien.asp (Stand: 24.3.2022)

[Rahmenrichtlinien IT] Rahmenrichtlinien der Schulen mit italienischer Unterrichtssprache: https://www.provinz.bz.it/formazione-lingue/scuola-italiana/sistema-scolastico/indicazioni-provinciali.asp (Stand: 24.3.2022)

Das neue Autonomiestatut (2019/1972): Geschichte und Kompetenzen; das Sonderstatut für die Region Trentino-Südtirol; das Staatsgesetz Nr. 118 vom 11. März 1972. https://www.landtag-bz.org/de/datenbanken-sammlungen/autonomiestatut.asp (Stand: 24.3.2022)

Fackelmann, J. (2000): Findefix: Wörterbuch für die Grundschule [1]. Mit lateinischer Ausgangsschrift. 3., überarbeitet Auflage. München.

Mayring, P. (2016): Einführung in die qualitative Sozialforschung. 6. Aufl. Weinheim/Basel

Nied Curcio, M. (2015): Spielen Wörterbücher bei der Sprachmittlung noch eine Rolle. In: Nied Curcio, M./Katelhön, P./Bašić, I. (Hg.): Sprachmittlung – Mediation – Mediazione linguistica: ein deutsch-italienischer Dialog. (= Sprachen lehren – Sprachen lernen). Berlin, S. 291–317.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Nkomo, D. (2015): Developing a dictionary culture through integrated dictionary pedagogy in the outer texts of South African school dictionaries: the case of Oxford Bilingual School Dictionary: IsiXhosa and English. In: Lexicography ASIALEX 2, S. 71–99.

Levy, M./Steel, C. (2015): Language learner perspectives on the functionality and use of electronic language dictionaries. In: ReCALL: the Journal of EUROCALL 27, S. 177–196.

Villa-Vigoni-Thesen 2018, hrsg. von einem Gremium aus Spezialistinnen aus der Wörterbuchforschung, der praktischen Lexikografie, dem Bereich Deutsch als Fremdsprache, der Italianistik, den Translationswissenschaften und der Empirischen Linguistik. DOI: https://hdl.handle.net/10863/9156 (Stand: 24.3.2022)

Kontaktinformationen

Andrea Abel
Institut für Angewandte Sprachforschung – Eurac Research
andrea.abel@eurac.edu

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Carolina Flinz/Sabrina Ballestracci

DAS LBC-WÖRTERBUCH: EINE ERSTE BENUTZERSTUDIE

Abstract This paper describes the results of an empirical investigation carried out within the project Lessico Multilingue dei Beni Culturali (LBC), whose aim is to create a multilingual online dictionary of the lexicon of the Italian artistic heritage. The dictionary, whose lexicographic process has already started, is intended for linguists and specialist translators as well as for professionals in the tourism sector and students of Foreign Languages and Literatures. The investigation conducted through a questionnaire submitted to undergraduate students at the University of Milan and at the University of Florence has a double aim: to research the habits in the use of lexicographic tools by possible users of the dictionary (Italian Learners of German Language), and to identify preferences regarding macro-, medio- and microstructural features of the future LBC-dictionary to realize a user-friendly tool. After a brief introduction on the state of the art of the survey in the field of Dictionary Users Studies, the article describes the questionnaire and the results obtained from the pilot study. A summary and a discussion on the future developments of the project conclude the work.

Keywords Dictionary use; LSP-Dictionary; Lexicon of Cultural Heritage

1. Einleitung[1]

Eine der zentralen Zielsetzungen der Forschungsgruppe *Lessico multilingue dei Beni Culturali* (LBC – ‚Mehrsprachige Lexik der Kulturgüter')[2] ist der Aufbau eines dynamischen und korpusbasierten mehrsprachigen Online-Fachwörterbuches zu den Florentiner Kulturgütern, das über das LBC-Wortinformationssystem frei verfügbar sein wird.[3] Das zu bearbeitende LBC-Wörterbuch richtet sich an ein heterogenes Publikum, insbesondere an WissenschaftlerInnen, (Fach-)ÜbersetzerInnen, StudentInnen sowie an Beschäftigte im Tourismusbereich, und zielt darauf ab, eine Lücke im lexikographischen Panorama zu schließen: Mehrsprachige lexikographische Ressourcen zum Wortschatz der Kunst und der Kulturgüter existieren noch nicht, sind jedoch für die Entwicklung von Sprach- und Fachkompetenzen sowie als Nachschlagewerk für Fachleute sehr wichtig (vgl. Farina/Garzaniti 2013; Farina 2016).

Primärquellen des Wörterbuches sind die in der ersten Arbeitsphase des Projekts zusammengestellten monolingualen LBC-Korpora in sechs Sprachen (Deutsch, Englisch, Französisch, Italienisch, Russisch, Spanisch).[4] Auf Basis der monolingualen Korpora wurden bereits erste Lemmalisten und eine Auswahl an Konkordanzen – Resultat einer feinkörnigen

[1] Der vorliegende Aufsatz wurde von den Autorinnen gemeinsam erarbeitet. Die einzelnen Kapitel lassen sich nicht individuell zuordnen. Die Anteile an der Urheberschaft betragen je 50%.

[2] Die Forschungseinheit *Lessico Multilingue dei Beni Culturali* wurde 2013 an der Universität Florenz gegründet. Weitere Kooperationen sind anschließend entstanden, darunter auch diejenige mit der Universität Mailand. Als lexikalisches Informationssystem wird die LBC-Plattform sowohl das Wörterbuch als auch die Korpora enthalten (vgl. https://www.lessicobeniculturali.net).

[3] Zum lexikographischen Prozess vgl. Farina/Flinz (2020).

[4] Diese Korpora, welche frei zugänglich sind und mit der Anwendung NoSketchEngine durchsucht werden können, können als Grundlage für vielfältige intra- und interlinguale Analysen verwendet werden (vgl. Ballestracci/Buffagni/Flinz 2020; Ballestracci 2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

qualitativen Analyse – fertiggestellt.[5] Da sich das Wörterbuch im Aufbau befindet (Storrer 2001, S. 65), ist es besonders wichtig, benutzergerechte Entscheidungen zu Makro-, Mikro- und Zugriffstrukturen zu treffen, um benutzerdefinierten Erwartungen bzw. Bedürfnissen entgegenzukommen. Zu diesem Zweck haben wir an der *Università degli Studi di Milano* und an der *Università degli Studi di Firenze* eine erste Umfrage mit potentiellen BenutzerInnen durchgeführt, deren Ergebnisse im vorliegenden Aufsatz vorgestellt werden. Dabei stützen wir uns einerseits auf existierende empirische Studien, deren ProbandInnen DaF-Lernende sind und die sich überdies zum Ziel setzen, Benutzergewohnheiten und -bedürfnisse zu erforschen, die auch mit dem Online-Medium verbunden sind (vgl. u. a. Balbiani 2022; Dominguez Vazquez 2013; Flinz 2014; Nied-Curcio 2016; Müller-Spitzer et al. 2018; Runte 2015), andererseits auf einen noch geringen Anteil von Untersuchungen, die in einem direkten Zusammenhang mit den Wörterbuchprojekten stehen (vgl. Flinz 2018; Meliss 2015).

Diese Pilotstudie, im Sinne eines *simultaneous feedback* (vgl. de Schryver/Prinsloo 2000), möchte zunächst die aktuellen Gewohnheiten und Bedürfnisse sowie hypothetischen Benutzungssituationen einer der avisierten Hauptbenutzergruppen des LBC-Wörterbuches rekonstruieren und analysieren: der Lernenden des Deutschen als Fremdsprache. Zum anderen steht unsere Untersuchung in einem direkten Zusammenhang mit dem LBC-Wörterbuchprojekt: Durch die Analyse sollen Desiderata der DaF-Lernenden konkret im Hinblick auf das zukünftige LBC-Wörterbuch bewertet werden, mit dem Ziel, sie in die Wörterbuchkonzeption einfließen zu lassen (Kap. 2). Dazu wird der Fokus im zentralen Teil der Arbeit (Kap. 3) besonders auf folgende Aspekte gelegt: 1) die bevorzugten Einstellungen der ProbandInnen gegenüber lexikographischen Online-Ressourcen und lexikalischen Online-Informationssystemen sowie unterschiedlichen Wörterbuchtypen; 2) die Handlungsmodalität, Handlungsfrequenz und Zweck der Suchanfrage; 3) die Makro-, Mikro-, Vernetzungs- und Zugriffstrukturen, die sie bei monolingualen und bilingualen allgemeinen Online-Wörterbüchern bevorzugen; 4) die Makro-, Mikro-, Vernetzungs- und Zugriffstrukturen, die sie bei einem Online-Fachwörterbuch der Kunst und Kulturgüter erwarten bzw. sich wünschen.

Die erhaltenen Ergebnisse können nicht nur wichtige Inputs für die zukünftige Gestaltung des LBC-Wörterbuches geben, sondern auch als Vergleichsparameter für Umfragen bei anderen durch das Projekt anvisierten Benutzergruppen (z. B. FachübersetzerInnen) dienen. Die Abbildung der Gewohnheiten unserer DaF-Lernenden ist darüber hinaus für eine bestimmte Benutzergruppe repräsentativ, so dass die Resultate können Anregungen für den Aufbau anderer lexikographischer DaF-Ressourcen bzw. Wörterbücher sowie Reflexionen in didaktischer Sicht bieten können (Kap. 4).

2. Umfragen in der Wörterbuchbenutzungsforschung

Umfragen sind eine der möglichen Methoden, die in der Wörterbuchbenutzungsforschung angewendet werden können[6]. Sie haben die längste Tradition (Taborek 2018, S. 208) und sind die häufigste verwendete Methode in der empirischen Sozialforschung (Müller-Spitzer 2018, S. 715).

[5] Überlegungen für das Deutsche finden sich in Buffagni/Flinz/Ballestracci (i. Ersch.). Für das Spanische ist die LBC-Liste mit ihren 2.000 Lemmata und 10.000 Konkordanzen bereits online unter http://lexicon.lessicobeniculturali.net/es/lemmario zugänglich.

[6] Für eine detailliertere Beschreibung der Methoden der empirischen Wörterbuchbenutzungsforschung vgl. u. a. Welker (2010); Engelberg/Lemnitzer (2009).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Umfragen können sowohl im Print-Format als auch im Online-Format mit unterschiedlichen Tools sowie in mündlicher Form durchgeführt werden (Engelberg/Lemnitzer 2009). Ihre Erstellung und Durchführung sind mit wenigen Komplikationen verbunden, obwohl die Tatsache, dass ProbandInnen nicht immer wahrheitsgemäß antworten, d. h. die Reaktivität der Methode, als möglicher Kritikpunkt gesehen werden kann (Müller-Spitzer 2016, S. 306). Trotz dieser Problematik (vgl. auch dazu Bergenholtz/Johnsen 2005; Tarp 2009) bleiben Umfragen, die sich auf allgemeine Daten oder Erfahrungen beziehen, eine weiterhin vertretbare und valide Methode (Müller-Spitzer 2018, S. 722), um Ergebnisse zu den überindividuellen Tendenzen von Wörterbuchbenutzern zu gewinnen, die für die Wörterbuchkonzeption und das Wörterbuchdesign hilfreich sein können.

Umfragen können zu unterschiedlichen Zwecken getätigt werden und die Spannbreite reicht von allgemeinen Untersuchungen zu Benutzergewohnheiten und -präferenzen bis zu einzelnen Wörterbuchtypen oder Wörterbüchern oder sogar zu Wörterbuchprojekten. In diesem letzten Fall können sie an unterschiedlichen Stellen des lexikographischen Prozesses durchgeführt werden, wie z. B. in der Vorbereitungsphase, nach Fertigstellung eines Onlinereleases oder zur Vorbereitung einer neuen Funktionalität.

Im deutsch-italienischen Raum kann ein steigendes Interesse für Wörterbuchbenutzungsstudien für den DaF-Unterricht wahrgenommen werden (Balbiani 2022; Nied Curcio 2011; Flinz 2014, 2014a; Mollica 2016; Müller-Spitzer et al. 2018). Untersuchungen unter DaF-Lernenden für Wörterbuchprojekte sind hingegen noch eine Rarität (Flinz 2018), auch wenn die BenutzerInnen als zentrale Größe bei der Planung und Erstellung von Wörterbüchern gesehen werden, da sie hergestellt werden, um als Hilfsmittel benutzt zu werden (Tarp 2008; Wiegand 1998, S. 259–260; Wiegand et al. 2010, S. 680; Müller-Spitzer 2016, S. 294).

Auch unsere Studie basiert auf einer Umfrage zur Wörterbuchbenutzung[7]. Die Umfrage besteht aus fünf Frageblöcken. Der erste Block enthält acht Fragen zu den soziolinguistischen Daten und zur Sprachbiographie der Studierenden (u. a. Alter, Geschlecht, Studiengang, Muttersprache, erlernte Sprachen und Sprachniveau). Die anderen vier Blöcke enthalten insgesamt 32 Fragen, die mit einem besonderen Fokus auf das LBC-Wörterbuchprojekt zum Ziel haben, Informationen zu den Tendenzen sowie Präferenzen der ProbandInnen in der Wörterbuchbenutzung zu erheben und abzubilden (vgl. Abbildung 1): Im ersten Block konzentrieren sich die Fragen auf die von den ProbandInnen bevorzugten Wörterbuchformate und lexikographischen Ressourcen (Fragen 1–9); im zweiten auf die Handlungsmodalität und -frequenz sowie zum Zweck der Suche (Fragen 10–14); im dritten auf die bevorzugten Makro-, Mikro-, Vernetzungs- und Zugriffstrukturen (Fragen 15–21); im vierten auf Erwartungen und Desiderata der Studierenden hinsichtlich eines Online-Fachwörterbuches der Kunst und Kulturgüter bzw. des LBC-Wörterbuches (Fragen 22–32).

[7] Das Design der Umfrage basiert auf der Terminologie in Wiegand (1998, S. 259 f.), Wiegand et al. (2010, S. 680) und Müller-Spitzer (2016, S. 294). Sie wurde in italienischer Sprache durchgeführt. Einige Fragen wurden vereinfacht, um die Verständigung zu erleichtern, da unsere Studierenden meistens keine Kurse zu lexikographischen Themen besucht haben.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

QUESTIONARIO

Età	☐17-21 ☐22-24 ☐25-29 ☐più di 30	
Sesso	☐M ☐ F ☐ N	
Lingua madre (LM)	☐Italiana ☐ _____	
Corso di Laurea		
Anno frequentato:	☐ 1. Anno (LT) ☐ 1. Anno (LM) ☐ 2. Anno (LT) ☐ 2. Anno (LM) ☐ 3. Anno (LT)	
Lingue straniere studiate (LS): inserire se L1, L2, L3	•Inglese: ☐L1 ☐ L2 ☐L3 ☐L4 •Francese: ☐ L1 ☐ L2 ☐L3 ☐L4 •Spagnolo: ☐L1 ☐ L2 ☐L3 ☐ L4 •**Tedesco: ☐ L1 ☐ L2 ☐L3 ☐L4** ☐ Altre.....: ☐L1 ☐ L2 ☐L3 ☐L4	☐ principianti ☐ progrediti ☐madrelingua ☐ principianti ☐ progrediti ☐ madrelingua ☐ principianti ☐ progrediti ☐ madrelingua ☐ principianti ☐ progrediti ☐ madrelingua ☐ principianti ☐ progrediti ☐ madrelingua
Certificazioni	•Inglese: ☐A1 ☐ A2 ☐ B1 ☐ B2 ☐ C1 ☐ C2 ☐ Altre ☐ non ho certificazioni •Francese: ☐A1 ☐ A2 ☐ B1 ☐ B2 ☐ C1 ☐ C2 ☐ Altre ☐ non ho certificazioni •Spagnolo: ☐A1 ☐ A2 ☐ B1 ☐ B2 ☐ C1 ☐ C2 ☐ Altre ☐ non ho certificazioni •**Tedesco**: ☐A1 ☐ A2 ☐ B1 ☐ B2 ☐ C1 ☐ C2 ☐ Altre ☐ non ho certificazioni ☐ Altre.....: ☐A1 ☐ A2 ☐ B1 ☐ B2 ☐ C1 ☐ C2 ☐ Altre ☐ non ho certificazioni	
Studio del Tedesco: indicare il numero di anni	☐ ..... Anni	
1) Quale **formato** di dizionario usi **abitualmente**? Se usi entrambi metti il numero 1 a quello che usi di più e 2 a quello che usi di meno. Perché? (indica almeno una motivazione, al massimo tre)	☐ dizionario cartaceo **Motivazioni:**	☐ dizionario online **Motivazioni:** ....
2) Quali sono a tuo avviso **gli aspetti positivi / negativi** dei due **formati**? (indica almeno **un** aspetto)	☐ aspetti positivi: .... ☐ aspetti negativi:	☐ aspetti positivi: .... ☐ aspetti negativi: ....
3) Quale **strumento lessicografico o informativo** usi abitualmente? Per ognuno indica anche la **frequenza**.	☐ dizionario ☐ corpora paralleli ☐ traduttore automatico ☐ corpora comparabili ☐ non conosco nessuno di questi strumenti	☐ spesso ☐ raramente ☐ mai ☐ spesso ☐ raramente ☐ mai ☐ spesso ☐ raramente ☐ mai ☐ spesso ☐ raramente ☐ mai
4) Quale **tipologia** di dizionario o usi abitualmente? Per ognuno indica anche la **frequenza**.	☐ monolingue ☐ bilingue / plurilingue	☐ spesso ☐ raramente ☐ mai ☐ spesso ☐ raramente ☐ mai
5) Quale **tipologia** di dizionario o usi abitualmente? Per ognuno indica anche la **frequenza**.	☐ generale ☐ specialistico ☐ altro	☐ spesso ☐ raramente ☐ mai ☐ spesso ☐ raramente ☐ mai ☐ spesso ☐ raramente ☐ mai
6) **Dizionari monolingui.** Per ognuno indica anche la **frequenza**.	☐ canoo.net ☐ Duden ☐ Elexiko ☐ Wiktionary ☐ DWDS-Wörterbuch ☐ Wortschatz-Portal Leipzig ☐ altri:	☐ spesso ☐ raramente ☐ mai ☐ spesso ☐ raramente ☐ mai ☐ spesso ☐ raramente ☐ mai ☐ spesso ☐ raramente ☐ mai ☐ spesso ☐ raramente ☐ mai ☐ spesso ☐ raramente ☐ mai ☐ spesso ☐ raramente ☐ mai

Abb. 1: Screenshot der ersten Seite der Umfrage

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

3. Ergebnisse der Umfrage

Die Umfrage wurde zu Beginn des Sommersemesters 2022 an der Universität Mailand und an der Universität Florenz durchgeführt. Die Umfrage ist Teil einer Studie, die im Sinne einer Trenddesignstudie[8] aktuell noch weitergeführt wird. Sie soll als Grundlage für weitere Studien dieser Art dienen, die zusätzliche Fremdsprachenlernende sowie andere BenutzerInnen-Gruppen involvieren soll. Im Folgenden werden die Ergebnisse der ersten Querschnittstudie vorgestellt, an der 53 ProbandInnen (43 weibliche und 10 männliche) teilgenommen haben. Die Teilnehmenden sind durchschnittlich in einem Alter zwischen 17 und 21 Jahren, besuchen das zweite bzw. das dritte Studienjahr des Bachelorstudienganges für fremdsprachige Literatur- und Sprachwissenschaft und studieren Deutsch als Fremdsprache. Sie sind italienische MuttersprachlerInnen, mit Ausnahme von zwei Studierenden, die als bilingual (deutsch-italienisch und spanisch-italienisch) SprecherInnen zu bezeichnen sind. Deutsch wird meistens als zweite oder dritte Fremdsprache nach Englisch bzw. Spanisch erlernt; eine kleine Gruppe von Teilnehmenden studiert auch Russisch (vier Testpersonen) oder nordische Sprachen (vier Testpersonen). 34 ProbandInnen haben ein fortgeschrittenes Niveau (Sprachniveau B2) und besuchen Deutschkurse seit mehr als fünf Jahren, während 19 AnfängerInnen (Sprachniveau A2) sind.

Im Folgenden werden wir die wichtigsten Ergebnisse der Umfrage vorstellen und diskutieren. Die Illustration erfolgt auf der Basis der am Anfang der Arbeit formulierten Forschungsfragen (vgl. Kap. 1) und besteht dementsprechend aus vier Blöcken (vgl. Kap. 2).

1) Der größte Teil der ProbandInnen (47) bevorzugt Internetwörterbücher (IWB), auch wenn ein Teil davon gerne noch beide Formate verwendet (19). Nur eine kleinere Gruppe benutzt ausschließlich Printwörterbücher (PWB) (4). Die positiven und negativen Aspekte, die für die jeweiligen Formate genannt werden, sind homogen (vgl. Tab. 1):

	+	–
IWB	schnellere Suche, höhere Erfolgsquote in der Suche, Aktualisierung des Produktes, Vorhandensein von Audiodateien, Speicherungsmöglichkeit von Recherchen, keine Kosten	weniger Beispiele, mehr Fehler, niedrigere Zuverlässigkeit
PWB	mehr Beispiele, Vollständigkeit, Zuverlässigkeit, Benutzung auch ohne Internet, Benutzung während der Prüfungen	Preis, längere Suchzeit, keine Aktualisierung des Produktes, Handlichkeit

Tab. 1: Auswahl der häufigsten genannten Vorteile/Nachteile von IWB und PWBf

Bemerkenswert ist, dass die Zuverlässigkeit trotz der Existenz von guten und wissenschaftlichen IWB noch mit Printwörterbüchern in Verbindung gebracht wird. Dieses Ergebnis kann dadurch erklärt werden, dass in Italien während der Prüfungen im universitären und schulischen Bereich meistens ausschließlich PWB erlaubt sind. Interessant ist auch, dass das Vorhandensein von weniger Beispielen und Kontexten (u. a. Phraseologismen) bei IWB – wie auch bei anderen empirischen Studien (Mollica 2017, 2020; Flinz 2021) – bemängelt wird.

[8] Vgl. mehrere Querschnittserhebungen zum gleichen Thema zu mehreren Zeitpunkten (vgl. Müller-Spitzer 2016, S. 299).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Das Wörterbuch zeigt sich weitgehend noch als die bevorzugte lexikographische Ressource und wird von der Mehrheit der Testpersonen (50) sehr häufig benutzt. Übersetzungstools werden als zweithäufigste verwendete Ressource genannt (42 ProbandInnen – 66% häufig). Nur ca. die Hälfte der ProbandInnen verwendet Parallelkorpora (26 ProbandInnen – 61% häufig), während nur eine kleine Zahl Vergleichskorpora zu Rate zieht (16 ProbandInnen – 50% häufig)[9]. Bezogen auf den Wörterbuchtyp werden sowohl einsprachige als auch zweisprachige/mehrsprachige Ressourcen seitens der Studierenden verwendet, mit einer leicht höheren Zahl des letzteren Typs (52 ProbandInnen vs. 40 ProbandInnen). Erheblich höher ist die Präferenz für allgemeine Sprachwörterbücher (fast 100%) gegenüber Fachwörterbüchern (FWB): Nur 21 ProbandInnen benutzen sie und die Handlungsfrequenz ist selten (81% selten). Interessant wäre es, die Begründung dafür zu erfahren sowie die Beantwortung der Frage, welche FWB am meistens konsultiert werden. Unter den einsprachigen Online-Ressourcen wird am meisten *Duden online* gebraucht (46 ProbandInnen), als zweites *Wiktionary* (25 ProbandInnen) und an dritter Stelle das *DWDS* (10 ProbandInnen). Das *Wortschatzportal Leipzig* sowie *elexiko* und *Canoo.net* werden kaum verwendet, was auch damit zusammenhängen könnte, dass sie den ProbandInnen gar nicht bekannt sind. Das Wörterbuch *Pons* ist das am häufigsten gebrauchte zweisprachige WB seitens der italienischen DaF-Lernenden (38 ProbandInnen – 63% häufig), gefolgt von *Langenscheidt* (21 ProbandInnen – 81% häufig), *Wiktionary* (19 ProbandInnen – 42 % häufig), das auch als zweisprachiges Wörterbuch gebraucht wird, um Äquivalenzbeziehungen zu identifizieren, und *Leo* (11 ProbandInnen – 81% häufig). *Dict.cc* wird nur von zwei Teilnehmenden benutzt. Automatische Übersetzungstools werden auch verwendet: Präferiert wird das Tool *Google Translator* (35 ProbandInnen – 40% häufig) gegenüber *DeepL* (nur 10 ProbandInnen – 80% häufig). Die Antworten zeigen jedoch auch, dass eine gewisse Zahl an Studierenden (12) solche Tools überhaupt nicht verwenden. Unter den Parallelkorpora wird am meisten *ContextReverso* gebraucht (48 ProbandInnen – 77% häufig), während *Linguee* nur von 6 ProbandInnen genannt wird.

2) Die interviewten Studierenden benutzen sowohl Laptops, PCs, Tablets und Handys für die Beantwortung ihrer lexikographischen Fragen (7 ProbandInnen benutzen sogar alle 4 Gerätetypen): An erster Stelle sind Smartphones (47 – 91% häufig) und Laptops (41 – 92% häufig) zu finden, gefolgt von PCs (36 – 78% häufig) und Tablets (12 – 58% häufig). Bezüglich der Häufigkeit der Handlung zeigt sich, dass 60% der ProbandInnen nach eigenen Aussagen mindestens einmal pro Tag die Ressourcen nutzt, während 36% zwei- bis dreimal in der Woche die Ressourcen verwendet. Nur 4% der Interviewten benutzt die Tools nur wenige Male im Monat. Mehr als die Hälfte (28) hat lexikographische Apps auf ihre Geräte heruntergeladen: *ContextReverso* (15), *Pons* (8), *Leo* (6), *Google Translator* (4), *Langenscheidt* (2) und *DeepL* (1). Ein Proband nennt auch die App von *Zanichelli*.

Unsere DaF-Lernenden benutzen lexikographische Ressourcen zu unterschiedlichen Zielsetzungen: meistens, um einzelne Lexeme oder Mehrwortverbindungen zu verstehen (52 ProbandInnen – 98% häufig), aber auch für die Übersetzung (45 ProbandInnen für die Übersetzung und 44 ProbandInnen für die Rückübersetzung); weniger, auch wenn mit einer hohen Zahl (37 ProbandInnen), um Texte in deutscher Sprache zu schreiben. Dieser letzte Punkt ist mit dem Sprachniveau der Teilnehmenden verbunden, die B2-Niveau besitzen, obwohl sie mehr als sieben Jahre Deutsch lernen. Bemerkenswert ist, dass 33 ProbandInnen alle vier Zielsetzungen angeben.

[9] Unsere Ergebnisse korrespondieren mit denen von Müller-Spitzer et al. (2018), da auch dort Wörterbücher von den Studierenden am häufigsten genutzt werden.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

3) Die Umfrage hat die Ergebnisse aus weiteren vergangenen Studien (vgl. u. a. Flinz 2014a) bestätigt, die hervorgehoben haben, dass Umtexte kaum oder selten seitens der WB-Benutzenden gelesen werden. Unsere ProbandInnen-Gruppe zeigt eine Präferenz für Abkürzungsverzeichnisse (34 Fälle), die höchstwahrscheinlich mit dem Benutzertyp zusammenhängt. Nutzungshinweise werden seltener gelesen (17 Fälle), die Einleitung noch seltener (7 – nur 15% der Fälle). Bezüglich der Vernetzungsstruktur verhalten sich die ProbandInnen unterschiedlich: Eine Hälfte hat vermerkt, dass alle Arten von Links (Strukturlinks, Inhaltslinks) wichtig sind; die zweite Hälfte bevorzugt eine „gut überdachte" Anzahl von Links, da die Gefahr besteht, vom Ziel der Suche abgelenkt zu werden. Für den Ausgangspunkt der Suche wird weiterhin die schriftformbasierte Suche angegeben, während die Suchhandlung unterschiedlich erfolgt: Alle Testpersonen bevorzugen die eingabebasierte Suche mit mehr Optionen (Vorschläge zur Vervollständigung der Zeichenkette, das Anzeigen von Lemmata die graphemisch ähnlich sind, Lemmatisierung der flektierten Form); 40% finden die indexbasierte Suche weiterhin sehr nützlich und 55% sind für eine filterbasierte Suche, insbesondere in Hinblick auf Wortbildungsprodukte und Wortkombinationen. Die Präferenz bezüglich der benötigten Angaben in der Mikrostruktur des Eintrages können aus der folgenden Grafik (Abb. 2) entnommen werden:

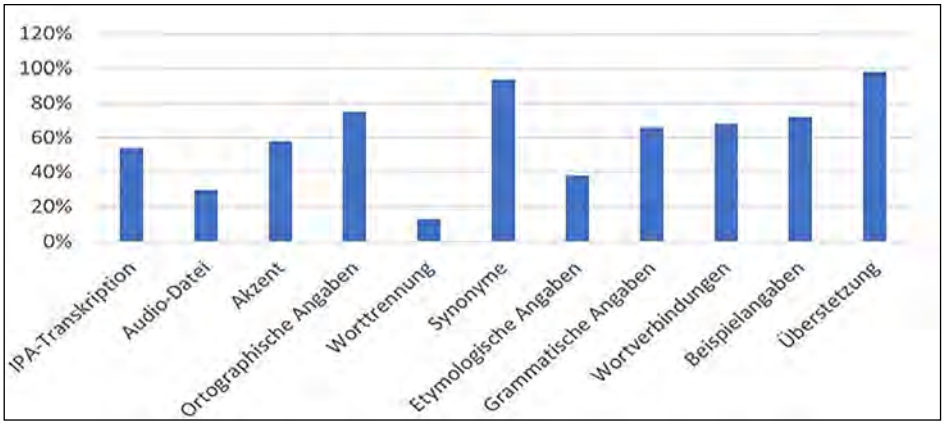


Abb. 2: Bevorzugte Angaben in der Mikrostruktur der Einträge

Die ProbandInnen beurteilen als unentbehrlich die Übersetzung und das Vorhandensein von paradigmatischen Angaben. Sehr hohe Prozentanteile haben auch orthographische, grammatische und syntagmatische Angaben. Weniger wichtig sind Informationen zur Akzentsetzung, zur Aussprache (IPA-Transkription und Audio-Datei), zu den etymologischen Angaben und zur Worttrennung.

4) Der letzte Teil des Fragebogens zielt darauf ab, Desiderata und Erwartungen abzubilden, die die untersuchte Benutzergruppe gegenüber dem LBC-Wörterbuch hat. Das Bild, das sich aus der Umfrage ergibt, ist für einige Aspekte eindeutiger, für andere variabler. Die meisten Studierenden würden ein solches Fachwörterbuch zu verschiedenen Zwecken benutzen, insbesondere für Tätigkeiten, mit denen sie in der Unterrichtspraxis konfrontiert werden: in erster Linie für die Übersetzung in die Fremd- sowie in die Muttersprache, in zweiter Linie für das Textverstehen und für die Textproduktion. Mehr als 80% der ProbandInnen hält nur einige der vorgeschlagenen Informationen für besonders nützlich: Wortkombinationen (z. B. *Verkündigung Maria*) und Wortvarianten (z. B. *Lorenzo il Magnifico*/*Lorenzo Magnifico*) stehen an erster Stelle. Diese Informationen empfinden sie als besonders hilfreich, um zu verstehen, wie ein Wort in einem bestimmten Kontext benutzt werden kann und wie es sich

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

im Laufe der Zeit entwickelt hat. Weniger nützlich finden sie das Vorhandensein von Eigennamen (z. B. *Santa Maria del Fiore*): Die angegebenen Begründungen dabei (z. B. 'Eigennamen werden normalerweise nicht übersetzt' oder 'Wörterbücher enthalten normalerweise keine Eigennamen') weisen darauf hin, dass die ProbandInnen an die konkreten Zwecke des Wörterbuches, wie etwa die Übersetzungspraxis, nicht auf seine mögliche Verwendung als Informationsquelle denken. Weitere Informationen, die fast alle ProbandInnen als unentbehrlich für ein Fachwörterbuch der Kunst und der Kulturgüter halten, sind immer noch diejenigen, die ihnen schnell helfen können, ein Wort zu verstehen bzw. zu übersetzen: das Äquivalent in der Fremd- oder Muttersprache, die Grundbedeutung, die Bedeutungsvarianten und die Gebrauchskontexte. Bezüglich der KWICs sind die Studierenden der Meinung, dass die Konkordanzen die Variation und den Spezifitätsgrad bei der kontextbezogenen Verwendung eines Wortes wiedergeben sollten. Weniger wichtig scheinen für die Lernenden andere Parameter zu sein, wie etwa die Länge der Zeilen sowie die Anordnung nach Sprachniveau. Als optional betrachten sie den Hinweis auf Synonyme, auf grammatische und etymologische Eigenschaften sowie auf die phonetische Transkription oder auf die lautliche Realisierung des Wortes.

Mit Bezug auf weitere befragte Eigenschaften (Informationen über die Primärquelle und entsprechenden Link zum Korpus, Verwendung von Farben) zeigen die Wünsche der Lernenden eine hohe Variabilität. Sehr oder ziemlich wichtig wird im Allgemeinen die Präsenz eines Hinweises auf die Primärquelle gehalten, von der auf die Spezifität sowie die Bedeutung des Eintrages geschlossen werden kann, obwohl eine große Anzahl von Lernenden diese Eigenschaft auch als nicht unentbehrlich sieht. Variabler ist die Einstellung gegenüber der Verwendung von Farben: Der Prozentsatz der Testpersonen, die diese Eigenschaft als (ziemlich) wichtig bewertet, gleicht dem Prozentsatz der Testpersonen, die sie als wenig oder kaum relevant sieht. Während einige Studierende erkennen, dass Farben zur besseren Benutzung des Wörterbuches, zur Erkennung der unterschiedlichen Wörterbuchsektionen und zur einer Memorisierung der Information beitragen können (eine Studierende weist auf die Wichtigkeit dieses Aspektes für Personen mit Lernschwierigkeiten hin), hat diese Eigenschaft für andere keine relevante Funktionalität.

Abschließend wurden die ProbandInnen um weitere Ratschläge/Hinweise gebeten: Die meisten derjenigen, die auf diese optionale Frage eine Antwort gaben, denken, dass ein Fachwörterbuch lediglich auf sehr technische und spezifische Informationen fokussiert sein sollte. Eine allzu elaborierte Mikrostruktur mit Hinweisen z. B. auf grammatische, etymologische, phonetische Eigenschaften, also auf Eigenschaften, die sie als Nebeninformationen betrachten, wird eher als verwirrend empfunden.

4. Fazit

In dieser Arbeit haben wir uns auf eine bestimmte avisierte Benutzergruppe des LBC-Wörterbuches konzentriert: auf DaF-Bachelorstudierende. Dabei handelt es sich um eine Benutzergruppe mit besonderen Bedürfnissen und Wünschen, die in den meisten Fällen konkreter Natur sind und mit den alltäglichen Tätigkeiten zusammenhängen, mit denen sie in der Unterrichtspraxis konfrontiert werden. Als unentbehrlich sehen sie alle diejenigen Eigenschaften eines Wörterbuches, die beim Verstehen oder bei der Übersetzung eines Wortes helfen können, d. h.: Angabe des Äquivalents, der Wortbedeutung, der Wortvarianten und des Kontextes. Weitere Informationen, die zum Verstehen der kulturellen Substanz und Wichtigkeit eines beliebigen Fachwortes beitragen können (z. B. Etymologie) oder zur

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Durchführung anderer konkreter Aktivitäten, wie beispielsweise die Produktion von fachlichen Texten, werden oft als sekundär bewertet.

Dieselbe Umfrage würde bei anderen Benutzergruppen sehr wahrscheinlich zu ganz anderen bzw. unterschiedlichen Ergebnissen führen. Schon Masterstudierende, die sich in ihrer alltäglichen Unterrichtspraxis mit der Entwicklung von Fachkompetenzen auseinandersetzen, würden vermutlich andere Antworten auf einige der gestellten Fragen geben und auch komplexere Tools präferieren, die sie bei der Interpretation und Produktion von Texten unterstützen. So würden vielleicht auch FachübersetzerInnen reagieren, denen z. B. die Kenntnis des kulturellen Umfeldes eines Wortes Übersetzungsstrategien suggerieren kann. Noch weitere Desiderata würden (Sprach-)WissenschaftlerInnen fordern, die sich mit der Fachlexik oder mit der Erforschung der Kunst und Kulturgüter beschäftigen.

Diese Pilotstudie hat zu Ergebnissen geführt, die in verschiedene Richtungen weisen. Einerseits können die hierdurch erhaltenen Daten als Vergleichsparameter für zukünftige Umfragen bei den anderen genannten Benutzergruppen auch von anderen Fremdsprachen verwendet werden. Andererseits hat sich die Studie als sehr hilfreich erwiesen, um Inputs für die mögliche Gestaltung des LBC-Wörterbuches zu bekommen. Die Benutzung eines Wörterbuches hängt sehr stark mit den Desiderata der einzelnen Benutzergruppen zusammen. Wörterbucheigenschaften, die die Bedürfnisse einer Gruppe befriedigen, können von einer anderen Benutzergruppe als irrelevant oder sogar störend empfunden werden. Die geplante Flexibilität der zukünftigen Online-Ressource möchte diesen Desiderata entgegenkommen. Darüber hinaus bietet uns diese Pilotstudie erste Erkenntnisse über die Einstellungen von jungen (DaF-)Studierenden gegenüber der Verwendung von Wörterbüchern sowie Anregungen in didaktischer Hinsicht. Bevorzugt sind von jungen Studierenden diejenigen Wörterbucheigenschaften, die den Lernenden eine unmittelbare Lösung konkreter Probleme bieten und die sie schnell und mit Erfolg bei ihren didaktischen Rezeptions- und Übersetzungstätigkeiten anwenden können (das Wörterbuch im Sinne eines konkreten Werkzeuges). Dass die Fachlexik der Kunst und Kulturgüter, wie diejenige anderer Fachsprachen, nicht lediglich eine Menge von isolierten Einzelwörtern ist, dass Wörter Wissensträger und -kondensate sind und dass diese aus ihrem Gebrauch hergeleitet werden können, dass fremdsprachliche Entsprechungen auch durch den Aufbau autonomer Recherchemethoden identifiziert werden können und dass neue Online-Instrumente mit Sprachreflexion verbunden sind – eine solche Bewusstheit zu vermitteln, ist die Aufgabe einer DaF-Didaktik, die mit Hilfe adäquater lexikographischer Ressourcen, Lernende zu motivierendem und somit auch erfolgreichem Spracherwerb führen kann.

Literatur und Online-Ressourcen

Balbiani, L. (2022): Studierende und Wörterbuchbenutzung im digitalen Zeitalter. In: Cantarini, S./Missaglia, F./Bertollo, S. (Hg.): Digitale Lehr-, Lern- und Forschungsressourcen für die deutsche Sprache. L'Analisi Linguistica e Letteraria 30 (2022/1), S. 107–124.

Ballestracci, S. (2022): Für die universitäre DaF-Didaktik sind sprachwissenschaftlich konzipierte Online-Korpusressourcen eine Ressource! In: Cantarini, S./Missaglia, F./Bertollo, S. (Hg.): Digitale Lehr-, Lern- und Forschungsressourcen für die deutsche Sprache. L'Analisi Linguistica e Letteraria 30 (2022/1), S. 173–192.

Ballestracci, S./Buffagni, C./Flinz, C. (2020): Il corpus LBC tedesco: costruzione e possibili applicazioni. In: Billero, R./Farina, A./Nicolás Martínez, M. C. (Hg.): I Corpora LBC. Informatica Umanistica per il Lessico dei Beni Culturali. Florenz, S. 55–75.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Buffagni, C./Flinz, C./Ballestracci, S. (i. Ersch.): Das deutsche LBC-Korpus: Provisorische Stichwortliste und Konkordanzen. In: Flinz, C./Buffagni, C./Ballestracci, S./Billero, R./Farina, A./Nicolás, Martínez, M. C. (Hg.): *Deutsche Lexik der Kunst auf der Basis des Korpus LBC* (*Lessico dei Beni Culturali*). Florenz.

Carpi, E./Pano Alamán, A./Billero, R./Farina, A./Nicolás Martínez M. C. (2021): Léxico español del arte documentado en el corpus LBC (Lessico dei Beni Culturali). Florenz. http://lexicon.lessicobeniculturali.net/es/lemmario (Stand: 23.3.2022).

De Schryver, G./Prinsloo, D. J. (2000): Dictionary-making process with 'Simultaneous Feedback' from the target users to the compilers. In: Heid, U./Evert, S./Lehmann, E./Rohrer, C. (Hg.): Proceedings of the Ninth EURALEX International Congress. Stuttgart, S. 197–209.

Domínguez Vázquez, M. J./Mirazo Balsa, M./Vidal Pérez, V. (2013): Wörterbuchbenutzung: Erwartungen und Bedürfnisse. Ergebnisse einer Umfrage bei Deutsch lernenden Hispanophonen. In: Domínguez Vázquez, M. J./Mirazo Balsa, M./Vidal Pérez, V. (Hg.): Trends in der deutsch-spanischen Lexikographie. Frankfurt a. M., S. 135–172.

Farina, A. (2016): Le portail lexicographique du Lessico plurilingue dei Beni Culturali, outil pour le professionnel, instrument de divulgation du savoir patrimonial et atelier didactique, PUBLIF@RUM, Bd. 24. https://www.publifarum.farum.it/index.php/publifarum/article/view/564 (Stand: 23.3.2022).

Farina, A./Flinz, C. (2020): LBC-Dictionary: a multilingual cultural heritage dictionary. Data collection and data preparation. In: Gavriilidou, Z./Mitsiaki, M./Asimakis, F. (Hg.): Lexicography for inclusion. Proceedings of the 19th EURALEX International Congress, 7–9 September 2021, Alexandroupolis. Bd. 1. Alexandroupolis, S. 371–379. https://euralex2020.gr/wp-content/uploads/2020/11/EURALEX2020_ProceedingsBook-p371-379.pdf (Stand: 23.3.2022).

Flinz, C. (2014): Wörterbuchbenutzung: Ergebnisse einer Umfrage bei italienischen DaF-Lernenden. In: Abel, A./Vettori, C./Ralli, N. (Hg.): Proceedings of the XVI EURALEX International Congress: The User in Focus, S. 213–224. www.euralex.org/proceedings-toc/euralex_2014 (Stand: 23.3.2022).

Flinz, C. (2014a): Wörterbuchbenutzung: Ergebnisse einer Umfrage unter Studenten der Tourismuswissenschaft. In: Mann, M. (Hg.): Digitale Lexikographie. Ein-und mehrsprachige elektronische Wörterbücher mit Deutsch: aktuelle Entwicklungen und Analysen. Hildesheim, S. 13–33.

Flinz, C. (2018): Der lexikographische Prozess bei Tourlex (ein deutsch-italienisches Fachwörterbuch zur Tourismussprache) für italienische DaF-Lerner. In: Klosa, A./Storrer, A./Taborek, J. (Hg.): Internetlexikographie und Sprachvermittlung. Jahrbuch Lexicographica. Berlin, S. 9–36.

Flinz, C. (2021): 'Wird eine Reise storniert oder annulliert?' Kollokationen und mehr oder weniger feste Wortverbindungen in deutsch-italienischen Online-Wörterbüchern. In: Mellado Blanco, C./Mollica, F./Schafroth, E. (Hg.): Kollokationen. theoretische, forschungspraktische und fremdsprachendidaktische Überlegungen. Frankfurt a. M., S. 69–92.

Engelberg, S./Lemnitzer, L. (2009): Lexikographie und Wörterbuchbenutzung. Tübingen.

Lessico multilingue dei Beni Culturali: www.lessicobeniculturali.net/de (Stand: 23.3.2022).

Meliss, M. (2015): Was suchen und finden Lerner des Deutschen als Fremdsprache in aktuellen Wörterbüchern? Auswertung einer Umfrage und Anforderungen an eine aktuelle DaF-Lernerlexikographie. In: InfoDaF 42 (2015/4), S. 401–431.

Mollica, F. (2017): Wörterbuchkritik und Wörterbuchbenutzungsforschung: Wie benutzerfreundlich ist die Registrierung von Kollokationen in ein- und zweisprachigen (Deutsch-Italienisch) Wörterbüchern? In: Bielińska, M./Schierholz, S. (Hg.): Wörterbuchkritik. Dictionary criticism. Berlin/Boston, S. 133–171.

Mollica, F. (2020): Funktionsverbgefüge in ein- und zweisprachigen Wörterbüchern (für das Sprachenpaar Deutsch-Italienisch) aus der Perspektive der DaF-Benutzer. In: De Knop, S./Hermann, M. (Hg.):

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Funktionsverbgefüge im Fokus. Theoretische, didaktische und kontrastive Perspektiven. Berlin, S. 137–178.

Müller-Spitzer, C. (2016): Wörterbuchbenutzungsforschung. In: Klosa, A./Müller-Spitzer, C. (Hg.): Internetlexikografie. Ein Kompendium. Berlin, S. 291–342.

Müller-Spitzer, C. et al. (2018): Correct hypotheses and careful reading are essential: results of an observational study on Learners using online language resources. In: Lexikos 28. Stellenbosch, South Africa, S. 287–315.

Nied Curcio, M. (2015): Wörterbuchbenutzung und Wortschatzerwerb. Werden im Zeitalter des Smartphones überhaupt noch Vokabeln gelernt? In: InfoDaF 42 (2015/5), S. 445–468.

Runte, M. (2015): Wie benutzen fortgeschrittene DaF-Lernende Wörterbücher? Eine Eye-Tracking-Studie zur Benutzung von Lernerwörterbüchern und ein Vorschlag zu deren Verbesserung. In: InfoDaF 5, S. 476–498.

Storrer, A. (2001): Digitale Wörterbücher als Hypertexte: Zur Nutzung des Hypertextkonzepts in der Lexikographie. In: Lemberg, I./Schröder, B./Storrer, A. (Hg.): Chancen und Perspektiven computergestützter Lexikographie. Hypertext, Internet und SGML/XML für die Produktion und Publikation digitaler Wörterbücher. (= Lexikographica. Series Maior 107). Tübingen, S. 53–69.

Welker, H. A. (2010): Dictionary use: a general survey of empirical studies. Brasilia. http://www.let.unb.br/hawelker/images/sto-ries/professores/documentos/dictionary_use_research.pdf (Standf: 23.3.2022).

Wiegand, H. E. (1998): Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung – zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie. Teilbd. 1. Berlin/New York.

Wiegand, H. E. et al. (2010): Wörterbuch zur Lexikographie und Wörterbuchforschung. Berlin/New York.

Kontaktinformationen

Carolina Flinz
Università degli Studi di Milano
carolina.flinz@unimi.it

Sabrina Ballestracci
Università degli Studi di Firenze
sabrina.ballestracci@unifi.it

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Zoe Gavriilidou/Evi Konstandinidou

THE EFFECT OF AN EXPLICIT AND INTEGRATED DICTIONARY AWARENESS INTERVENTION PROGRAM ON DICTIONARY USE STRATEGIES

Abstract There is a growing interest in pedagogical lexicography, and more specifically in the study of dictionary users' abilities and strategies (Prichard 2008; Gavriilidou 2010, 2011; Gavriilidou/Mavrommatidou/Markos 2020; Gavriilidou/Konstantinidou 2021; Chatjipapa et al. 2020). The purpose of this presentation is to investigate dictionary use strategy and the effect of an explicit and integrated dictionary awareness intervention program on upper elementary pupils' dictionary use strategies according to gender and type of school. A total of 150 students from mainstream and intercultural schools, aged 10–12 years old, participated in the study. Data were collected before and after the intervention through the Strategy Inventory for Dictionary Use (SIDU) (Gavriilidou 2013). The results showed a significant effect of the intervention program on Dictionary Use Strategies employed by the experimental group and support the claim that increased dictionary use can be the outcome of explicit strategy instruction. In addition, the effective application of the program suggests that a direct and clear presentation of DUS is likely to be more successful than an implicit presentation. The present study contributes to the discussion concerning both the 'teachability' of dictionary use strategies and skills and the effective forms of intervention programs raising dictionary use awareness and culture.

Keywords Dictionary use strategies; explicit and integrated intervention program; dictionary culture; pedagogical lexicography

1. Introduction

Dictionary use strategies (DUS) are 'techniques' used by the effective dictionary user, in order to decide whether to use or not an appropriate type of dictionary and make a quick and successful search in it (Gavriilidou 2013). The author classifies DUS for paper dictionaries in four categories: 1) Dictionary awareness strategies which refer to the critical awareness of the value and shortcomings of the dictionary that lead to the decision to use a dictionary in order to resolve a specific problem encountered during learning inside or outside the classroom, 2) Dictionary selection strategies which allow the choice of an appropriate dictionary depending on the problem to be solved and guarantee the familiarity with one's own dictionary, 3) Lemmatization strategies, which help dictionary users find the citation form of inflected forms found in the text by relying on morphological indices (stems, prefixes, suffixes, inflectional morphemes) of the unknown word they come across in the/a text in order to make hypotheses about the look-up form of that word. Lemmatization strategies also include skills in alphabetical sequencing, otherwise lemmatization is not possible, and 4) Look-up strategies, which control and facilitate the localization of the correct section of the entry where different meanings of the same polysemous word form are included. Gavriilidou/Mavrommatidou/Markos (2020) propose the following DUS for digital dictionary use: 1) strategies familiarizing with different types of electronic dictionaries and the conditions of their use; 2) strategies for lemmatization and acquaintance with dictionary conventions; 3) navigation strategies; and 4) look-up strategies in new electronic environments.

Depending on the type of processing involved, these strategies can be further classified into metacognitive, cognitive, memory or compensating. Metacognitive DUS include self-man-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

agement, self-monitoring, self-reflection, decision making, planning, etc. and can be applied in receptive or productive dictionary use for conflict resolution or evaluating dictionary use success. They raise dictionary users' awareness of what they are doing and help them setting look up goals and deploying alternative plans when the goals are not met. Cognitive DUS include inferencing or alphabetization. Memory DUS include use of mnemonics to remember the word to be searched. Finally, compensation DUS, such as paying attention to headwords, signposts or example sentences enable dictionary users to better navigate in the dictionary and are intended to make up for inadequate information or skills.

A growing body of research has focused on the close relationship between dictionary use strategies and effective dictionary use (Chatjipapa et al. 2020; Gavriilidou/Mavrommatidou/Markos 2020), while Gavriilidou/Konstantinidou (2021, p. 735) showed that DUS are teachable. The authors also highlighted that

> strategic dictionary instruction should be an integral part of language education (first, second, foreign or heritage), since it helps students acquire dictionary culture, gain greater proficiency and confidence in dictionary use, and self-awareness about when and how we chose to use a dictionary in an autonomous way. (Ibid.)

While the literature on training reference skills and DUS is not overwhelming, there are already some useful findings focusing on the need to teach how to (strategically) use a dictionary effectively (Campoy-Cubillo 2002; Carduner 2003; Herbst/Stein 1987; Krieger/Müller 2017; Lew/Galas 2008; Zingano Kuhn 2019). Furthermore, Walz (1990) and Bishop (2000) are among the very few researchers who designed learning activities for training students how to use a dictionary.

Previous research has also highlighted that two crucial questions have to be taken into consideration when designing a syllabus or an intervention program for training DUS: the explicitness of purpose while teaching and the effectiveness of integrating strategy instruction into language class. However, no previous research investigated so far the impact of specific characteristics of an intervention program (such as explicitness of purpose and integration in the language course) nor the effect of variables such as gender, school type, multilingualism or dictionary use at home on DUS.

To bridge this gap in previous literature, this paper reports findings from a quantitative study conducted in Greek mainstream and intercultural schools for investigating the effect of an explicit and integrated dictionary awareness intervention program on upper elementary pupils' dictionary use strategies according to gender and type of school. The intervention was held within the class of Greek Language teaching and was based on the school dictionary distributed to all pupils by the Greek Ministry of Education. The research questions underlying this research were the following:

RQ1: What is the frequency of self-perceived DUSs of the sample and the individual dictionary use strategies that participants report they use the most/the least during digital dictionary consultation? Considering previous research (Chadjipapa et al. 2020) we expect a moderate overall strategy use and low to moderate strategy use to the four different types of digital dictionary use.

RQ2: What is the effect of the intervention program on DUS by gender and type of school?

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

This study extends previous research by offering additional arguments about the importance of teaching dictionary use skills and strategies and the teachability of DUS. It also offers useful insights about parameters affecting DUS.

2. Study

2.1 Participants

The sample consisted of 150 students with approximately equal numbers of males (49,3%) and females (50,7%). The participants attended in two different types of schools (mainstream and intercultural) in two Greek cities (Komotini and Ierapetra of Crete) and they were selected using convenience sampling. The participants attended the 5th and 6th grade of elementary school (upper elementary). The students were divided into two groups as follows: the control group consisted of two classes of grade 6 and one class of grade 5 with 25 students each (total of 75) and the experimental group consisted of two classes of grade 5 and one class of grade 6 with 25 students each (total of 75). The students in both groups participated in the diagnostic (pre-test) and evaluative (post-test) tests at the same time periods, but only the students in the experimental group participated in the teaching intervention activities. In terms of gender, the students are almost equally distributed in each group (Table 1).

Group	**Gender**	***n***	**%**
Experimental Group	Male	38	25,3
	Female	37	24,7
Control Group	Male	36	24,0
	Female	39	26,0
Total		**150**	**100**

Table 1: Distribution of students in the two groups by gender

2.2 Procedures and instrumentation

A quasi-experimental research method with a "pre-test-post-test control-group design" was adopted in the present study. The study was carried out in three stages. In the first stage, all the participants filled in the Strategy Inventory for Dictionary Use (SIDU) (Gavriilidou 2013) which is a valid and reliable self-report tool for the strategic use of dictionary. It consists of 36 five-point Likert-scale items, ranging from 1 (= never or almost never true of me) to 5 (= always true of me), belonging to four different subscales: (1) Dictionary awareness strategies; (2) Dictionary selection strategies; (3) Lemmatization strategies; and (4) Look-up strategies. At the end of the questionnaire there was an additional appendix that provides personal information on students' profiles such as gender, type of school, multilingualism and dictionary use at home. Cronbach's Alpha coefficient for the overall instrument was .93 suggesting an excellent degree of internal consistency. The value of the Alpha coefficient was: a) .87 for dictionary use awareness skills; b) .77 for dictionary selection strategies; c) .82 for strategies used in lemmatization and acquaintance with dictionary conventions; and finally d) .84 for look-up strategies. These values indicate a high degree of internal con-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

sistency in the overall instrument and all sub-scales, showing that the instrument provides internally consistent scores.

In the second phase, strategy-based training was carried out with the experimental group, while the control group received the standard FL instruction. The intervention program was applied to the students in the experimental group for a period of 4 weeks (2 hours per day) and finally, after the completion of the program, the measurement of the frequency of use of the strategies was repeated in the same time periods in both experimental and control groups.

In the final stage, which followed the completion of the treatment, strategy use was measured for both groups with the same instrument. Cronbach's Alpha coefficient for the overall instrument at the second measurement was .96 suggesting an excel-lent degree of internal consistency. The value of the Alpha coefficient was: a) .93 for dictionary use awareness skills; b) .85 for dictionary selection strategies; c) .87 for strategies used in lemmatization and acquaintance with dictionary conventions; and finally d) .86 for look-up strategies. The frequency of strategy use overall and for each of the strategy categories represents the dependent variable, expected to be influenced by the independent variables, which are the following: the intervention (experimental and control group) and the measurement (before and after the intervention).

2.3 The intervention program

The intervention program includes 12 units of targeted paper DUS instruction for pupils attending the two classes of upper elementary schools in Greece. Each unit corresponds to and is closely connected to a different chapter of the school textbook for teaching Greek as L1. The program may be conducted over a minimum of a 4-week period. However, the duration may be extended depending on the classroom needs, level and interest. The specific intervention includes activities that promote dictionary use strategies, which are listed in the SIDU and follows the principles of strategy-based instruction. Strategy-based instruction (SBI) enables learners to take an active role in the learning process by helping them to monitor and evaluate the way they learn. It is explicit and integrated, since the students learn the way, the reasons and the instances under which they can use the appropriate dictionary applying the suitable strategies during its implementation, and enables the learners to correct themselves and their mistakes during the learning process. It adopts differentiated learning and proposes adapted activities in order to respond to the needs of users with disabilities (learning difficulties, blindness, etc.). It is based on the textbooks and school dictionary distributed free of charge by the Greek Ministry of Education in Greek upper elementary schools, and finally, it clearly states the learning outcomes of each activity (for a detailed presentation of the intervention program see Gavriilidou/Konstantinidou 2021).

3. Statistics

Data elicited from students' responses to the SIDU questionnaire were analysed with SPSS version 23. Descriptive statistics (means and standard deviations) were calculated, in order to investigate the central tendency and dispersion of the student answers to the SIDU items. Furthermore, a two-way repeated measures ANOVA with group (experimental, control) as a between-subject factor and time (pre-test, post-test) as a within-subject factor was used to investigate the effect of the intervention on the frequency of strategy use between groups.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Finally, a post-hoc Bonferroni test was carried out where significant p-values (< 0.05) were found to determine which groups were significantly different. The level of statistical significance was set at $p < 0.05$. To determine the effect of gender and school type on the frequency of use of dictionary strategies by the sample students, a one-factor analysis of variance (One-way ANOVA) was performed. The significance level of the statistical tests was set at $\alpha = 0.05$.

4. Results

4.1 Descriptive statistics

The overall mean score of dictionary strategy use was found to reflect a moderate level of use (M = 2.86, SD = 0.76). Moreover, students reported low to moderate strategy use with regard to the four individual types of strategies, with look-up (LU) strategies having the highest mean score (M = 3.28, SD = 0.90), followed by lemmatization (LM) strategies (M = 3.01, SD = 0.98), dictionary selection (DS) (M = 2.77, SD = 0.90), and dictionary awareness (DA) strategies (M = 2.58, SD = 0.76), which were the least used strategies.

Tables 2 and 3 offer respectively an overview of the most least frequently reported strategies.

Strategy type	Mean	SD
When I look up a word, I constantly bear it in my mind during the search. **LU**	3,66	1,263
Before I buy a dictionary, I know the reason why I need it. **DS**	3,63	1,397
When I look up a word, I bear in mind its initial letter and then I search where I believe this initial letter is in the dictionary. **LU**	3,45	1,383
When I come across an unknown word in a text, I try to think in what form I should look it up in the dictionary.	3,41	1,457
When I look up a word beginning with E, I search in the first quarter pages as E is one of the first letters of the alphabet	3,41	1,381

Table 2: The most frequently reported strategies

Strategy type	Mean	SD
I know what an etymological dictionary is and what it is used for	2,37	1,308
I use a dictionary when I read a text	2,3	1,273
I use a dictionary to find antonyms	2,27	1,085
I use a dictionary to find the syntax of a word	2,19	1,191
I know what a terminology dictionary is and what it is used for	2,09	1,212

Table 3: The least frequently reported strategies

4.2 The effect of the intervention program on DUS

Means (and standard deviations) of the students' use of DUS in total for each group before and after the intervention are presented in table 4. The application of 2 x 2 ANOVA showed

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

that the two groups (experimental and control) were not statistically significantly different before the intervention in terms of the frequency of using the strategies overall (Mean Difference = 0.426, p = 0.07). Thus, the two groups can be considered equivalent before the implementation of the syllabus. The interaction between group and measurement was found to be statistically significant (F(1, 148) = 35.997, p< 0.001, η2 = 0.196). The value of η2 indicates that 19.6% of the variability in the use of strategies overall can be attributed to the statistical effect of group and measurement, which corresponds to a large effect size. In particular, before the intervention, students in the control group report low to moderate use of strategies overall and this frequency does not change significantly after the intervention (Mean Difference = 0.074, p = 0.298). In contrast, after the intervention, students in the experimental group state that they use the strategies to a significantly greater degree overall (M.D. = 0.676, p< 0.001).

Group	Before		After	
	M	S.D.	M	S.D.
Experimental Group (*n* = 75)	3,07	0,71	3,75	0,59
Control Group (*n* = 75)	2,65	0,66	2,72	0,62

Table 4: The effect of the intervention program on the level of dictionary use strategies

4.3 The effect of the intervention programme on DUS by gender and type of school

The means (and standard deviations) of the use of vocabulary strategies overall by boys and girls in each group before and after the intervention are presented in Table 5. The 2 x 2 x 2 ANOVA showed that the interaction between group, gender and measurement was not found to be statistically significant (F(1, 146) = 0.862, p = 0.355). Given that the interaction between group and measurement is statistically significant we conclude that the intervention had a statistically significant effect on both boys and girls in the experimental group.

Group		Before		After	
	Gender	M	SD	M	SD
Experimental Group (*n* = 75)	Male (*n* = 38)	2,77	0,69	3,62	0,59
	Female (*n* = 37)	3,38	0,59	3,88	0,58
Control Group (*n* = 75)	Male (*n* = 36)	2,48	0,73	2,65	0,70
	Female (*n* = 39)	2,79	0,55	2,78	0,5

Table 5: The effect of the intervention program on DUS according to gender

The means (and standard deviations) of the use of vocabulary strategies in general and intercultural schools in each group before and after the intervention are presented in Table 6. The 2 x 2 x 2 ANOVA showed that the interaction between group, school type and measurement was not found to be statistically significant (F(1, 146) = 0.067, p= 0.797). Given that the interaction between group and measurement is statistically significant we conclude that the intervention had a statistically significant effect for both general school students and those in the intercultural school in the experimental group.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Group	Type of school	Before		After	
		M	SD	M	SD
Experimental Group (n = 75)	General (n = 50)	2,97	0,67	3,77	0,58
	Intercultural (n = 25)	3,27	0,77	3,70	0,63
Control Group (n = 75)	General (n = 25)	2,53	0,64	2,81	0,53
	Intercultural (n = 50)	2,70	0,66	2,67	0,66

Table 6: The effect of the intervention program on DUS according to school type

5. Discussion and conclusions

The purpose of the present study was to investigate strategic dictionary use and the effect of an explicit and integrated dictionary awareness intervention program on upper elementary pupils' dictionary use strategies according to gender and type of school.

It was found that students attending Greek schools reported moderate overall use of DUSs and a more frequent use of look-up and lemmatization strategies. This result is in line with previous research (Chatjipapa et al. 2020; Gavriilidou/Mavrommatidou/Markos 2020) indicating that school-aged students in Greece consider that they use DUSs to some degree, and also that they are more familiar with look-up and lemmatization strategies. This finding suggests that more needs to be done in order to raise the moderate use of all types of DUS, cultivate a dictionary culture among elementary and secondary pupils and increase the awareness of the benefits of dictionary use and its potential in improving students' lexical knowledge. Of course, teachers' staff development through constant in-service training is necessary in order to gain expertise and be able to systematically incorporate DUS into the Greek educational setting.

The results also showed a significant effect of the intervention program on DUS employed by the experimental group. This finding provides additional support to the 'teachability' of dictionary use strategies and skills; It also suggests that the effective forms of intervention programs may raise dictionary use awareness and culture and support the claim that increased dictionary use can be the outcome of explicit strategy instruction. In addition, the effective application of the program suggests that a direct and clear presentation of DUS is likely to be more successful than an implicit presentation.

Furthermore, a statistically significant interaction was found between "group" (experimental vs. control) and "measurement" (pre-measurement vs. post-measurement). No statistically significant interaction was found between 'group', 'gender' and 'measurement', suggesting that both girls and boys of the experimental group increased their scores in DUS. This finding offers further support to Chatjipapa et al. (2020) who maintained that gender is not a strong predictor of DUSs. Moreover, it is in line with previous research that indicated that, in upper elementary, male and female students use DUSs equally (Chatjipapa et al. 2020).

Similarly, no statistically significant interaction was found between 'group', 'type of school' and 'measurement'. Given that a statistically significant interaction was found between "group" and "measurement", the above finding suggests that pupils of the experimental group attending both mainstream and intercultural schools benefited equally from the intervention program and that the intervention is a strong predictor of DUS.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

6. Limitations

There are certain restrictions in the present study. First of all, the study was based on a quantitative research design that included a questionnaire survey. The combination of quantitative and qualitative methods could lead to more reliable conclusions. In addition, the study was based in a convenient sample, so the results should be interpreted with caution. Finally, the post-test was conducted shortly after the intervention and may have only measured short-term results in dictionary use strategies.

References

Bishop, G. (2000): Developing learner strategies in the use of dictionaries as a productive language learning tool. In: The Language Learning Journal 22 (1), pp. 58–62.

Campoy-Cubillo, M. C. (2015): Assessing dictionary skills. In: Lexicography Asialex 2 (1), pp. 119–141.

Carduner, J. (2003): Productive dictionary skills training: what do language learners find useful? In: Language Learning Journal 28, pp. 70–76.

Chadjipapa, E./Gavriilidou, Z./Markos, A./Mylonopoulos, A. (2020): The effect of gender and educational level on dictionary use strategies adopted by upper-elementary and lower-secondary students attending Greek Schools. In: International Journal of Lexicography 33 (4), pp. 443–462.

Gavriilidou, Z. (2010): Profiling Greek adult dictionary users. In: Studies in Greek Linguistics 31, pp. 166–172.

Gavriilidou, Z. (2011): Users' abilities and performance in dictionary look up. In: Lavidas, N. (ed.): Selected papers of the 20th International Symposium of Theoretical and Applied Linguistics. Thessaloniki, pp. 41–51.

Gavriilidou, Z. (2013): Development and validation of the Strategy Inventory for Dictionary Use (S.I.D.U). In: International Journal of Lexicography 22 (2), pp. 135–154.

Gavriilidou, Z./Konstantinidou, E. (2021): The design of an explicit and integrated intervention program for pupils aged 10–12 with the aim to promote dictionary culture and strategies. In: Gavriilidou, Z./Mitis, L./Kiosses, S. (eds.): Proceedings of the XIX Euralex Congress: Lexicography for Inclusion. Vol. 2. Komotini, pp. 735–745. https://euralex2020.gr/wp-content/uploads/2021/09/Pages-from-EURALEX2021_ProceedingsBook-Vol2-p735-745.pdf.

Gavriilidou, Z./Mavrommatidou, S./Markos, A. (2020): The effect of gender, age and career orientation on digital dictionary use strategies. In: International Journal of Research Studies in Education 9 (6), pp. 63–76. https://doi.org/10.5861/ijrse.2020.5046.

Herbst, T./Stein, G. (1987): Dictionary-using skills: a plea for a new orientation in language teaching. In: Cowie, A. P. (ed.): The dictionary and the language learner. Papers from the EURALEX seminar at the University of Leeds, 1–3 April 1985. Tübingen, pp. 115–127.

Krieger, M. G./Müller, A. F. (2017): Caderno Interativo: atividades com o dicionário. Rio de Janeiro.

Lew, R./Galas K. (2008): Can dictionary skills be taught? The effectiveness of lexicographic training for primary-school-level Polish learners of English. In: Bernal, E./DeCesaris, J. (eds.): Proceedings of the XIII EURALEX International Congress. Barcelona, pp. 1273–1285.

Macaro, E. (2001): Learning strategies in foreign and second language classes. London.

Prichard, C. (2008): Evaluating L2 readers' vocabulary strategies and dictionary use. In: Reading in a Foreign Language 20 (2), pp. 216–231.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Walz, J. (1990): The dictionary as a secondary source in language learning. In: The French Review 64 (1), pp. 79–94.

Zingano Kuhn, T. (2019): State-of-the-art on monolingual lexicography for Brazil (Brazilian Portuguese). In: Slovenščina 2.0: empirical, applied and interdisciplinary research 7 (1), pp. 98–112. DOI: http://dx.doi.org/10.4312/slo2.0.2019.1.98-112.

Contact information

Zoe Gavriilidou
zoegab@otenet.gr
Greece Democritus University of Thrace

Evi Konstandinidou
evi1990@hotmail.gr
Greece Democritus University of Thrace

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Theresa Kruse/Ulrich Heid

LEARNING FROM STUDENTS

On the design and usability of an e-dictionary of mathematical graph theory

Abstract We created a prototype of an electronic dictionary for the mathematical domain of graph theory. We evaluate our prototype and compare its effectiveness in task-based tests with that of Wikipedia. Our dictionary is based on a corpus; the terms and their definitions were automatically extracted and annotated by experts (cf. Kruse/Heid 2020). The dictionary is bilingual, covering German and English; it gives equivalents, definitions and semantically related terms. For the implementation of the dictionary, we used LexO (Bellandi et al. 2017). The target group of the dictionary are students of mathematics who attend lectures in German and work with English resources. We carried out tests to understand which items the students search for when they work on graph-theoretical tasks. We ran the same test twice, with comparable student groups, either allowing Wikipedia as an information source or our dictionary. The dictionary seems to be especially helpful for students who already have a vague idea of a term because they can use the resource to check if their idea is right.

Keywords Dictionary use; specialised languages; terminology; terminography; mathematics

1. Introduction

Wikipedia "will eliminate the market for traditionally conceived specialised online dictionaries, which will disappear as commercial products if they cannot offer anything to justify their market price," says Fuertes-Olivera (2016). With this in mind, we want to learn more about the demands users have for specialised online dictionaries. The field of our study is the domain of graph theory and we created a prototype of an electronic dictionary on the subject which we evaluate with students (N=113). The participants are given ten tasks that can be solved with the help of our prototype and we asked them to comment on their working procedures. For comparison, a second group with the same tasks was instructed to complete them using Wikipedia (N=182). In this paper, we discuss the comments of both groups and focus on microstructure and access structure. The method is similar to the user-centered design approach which has previously been used to assess the usability of dictionaries, also motivated by the function theory (Tarp 2008). In particular, we would like to answer the following research questions:

- How beneficial is our dictionary for solving language-based mathematics tasks?
- What can we learn from student feedback about challenges, search workflows, efficiency and user satisfaction with a view to the development of future terminological e-dictionaries?

The findings may be applied to terminological dictionaries in other domains and provide additional insights into what terminological dictionaries have to offer to compete with Wikipedia.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

2. Related work

An increasing body of literature on the usability of electronic dictionaries has emerged in recent years. We only cite some exemplary work. A general review of the status around 2010 to 2015 is provided by Töpel (2014) and Nesi (2015); for a comparison between printed and electronic dictionaries see Dziemianko (2012). The majority of the publications focus on general language learner dictionaries, only a few on the acquisition of terminology. We only cite empirical results that discuss the design of e-dictionaries and how users interact with them; pure log-file analyses, studies on pocket dictionaries and meta-surveys are not mentioned.

2.1 Microstructure

The microstructure of e-dictionaries has been the subject of several user-oriented analyses. They mainly examine which items users would like to have included, which elements they use and how these items should be presented. Laufer/Hill (2000, p. 74) show that "even though a variety of dictionary information was available, most students opted for definitions, translations, or both". Li/Xu (2015) confirm this result for definitions, but in their study examples are rated high as well. We can conclude that users mostly demand definitions, translations and examples, depending on the particular use situation. Khairiah (2021) investigates what users expect from e-dictionary definitions: "The respondents consider that the information category of taxonomy/scientific name and characteristics are the most significant in a definition." Concerning the layout of the microstructure, Dziemianko (2015) shows that functional labels in colour significantly increase the speed and effectiveness of online dictionary searches over those in black and white. Chan (2014, p. 39) shows:

> The layout of a dictionary entry is also an important factor determining learners' preferences. A dictionary entry may be excessively long and may contain much more information than what a learner wants to get. The situation will be compounded if a target word is polysemous. [...] The use of highlighting features will of course help focus learners' attention, but a clear and systematic layout should help facilitate information search. As evidenced from the results of the study, learners prefer dictionary information which is clearly presented.

2.2 Role of illustrations

The role of illustrations in e-dictionaries is another area of research. Based on voluntary disclosure and an eye-tracking study, Kemmer (2014) finds that users consult pictures and text equally, with a slight preference for text. Lew et al. (2018, p. 73) confirm this and find "no evidence that the presence of pictorial illustrations leads to the neglect of the verbal definition." However, in the instance of graph theory, where certain concepts have typical graphical representations this might not be the case.

2.3 Search strategies

Another significant aspect is how users interact with electronic dictionaries and in particular which usage patterns appear. Laufer/Levitzky-Aviad (2006, p. 151) identify "translation, definition(s) and example(s)" as the most typical pattern for a bilingual dictionary, closely followed by "translation only". Aust et al. (1993) measure efficiency by consultations per

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

minute and compare printed and digitized dictionaries, finding that digital and bilingual (unlike monolingual) dictionaries have superior efficiency. Laufer/Hill (2000, pp. 68/77) conclude:

> The number of times the word is looked up during a learning session bears almost no relation to its retention. We postulated, albeit cautiously, that what matters is greater attention during the lookup rather than the number of lookups.

Chen (2017) confirm these results. Heid/Zimmermann (2012) compare a user interface that resembles a search engine to one that allows users to specify their needs in much detail:

> A comment made by most participants to the study was that they had needed some time to get acquainted with the profile-based dictionaries; this shows that these are not exactly conformant to users' a priori expectations. Our users however also noted that once the principles had been understood, the profile-based dictionaries were indeed more effective and efficient than the one-shot dictionary.

Lorentzen/Theilgaard (2012, pp. 654 f.) confirm the desire for a search engine-like user interface. Boonmoh (2021) let users freely choose between a translation tool and an e-dictionary and all of them used Google Translate in addition to other resources.

2.4 Method: work with user comments

Hult (2014) analyses and classifies dictionary users' comments into five categories: criticism, suggestions for improvement, praise, other and nonsense. If a comment fit into several categories, it was divided. She also examines how often different parts of the microstructure are mentioned in the comments and if so whether positively or negatively. Miller et al. (2017) analyse user comments qualitatively and discover audio, easy to understand, easy to use, example sentences, free access and pictures as popular features while font size, offline use, number of entries and example sentences leave room for improvement. He (2017) uses a grounded theory approach to classify comments and finds the following categories, with sub-categories listed in parentheses: comprehension (definition, collocation, illustration), production (thesaurus, example, pragmatic information, usage notes, morphology), information accessibility (headword selection, neologism, pronunciation, encyclopedic information). Farina et al. (2019) use think-aloud protocols to examine how the dictionaries in use could be improved. They find that users choose dictionaries based on search engine rankings, that part of speech and lemmatisation are neglected during the search process and that users desire a sufficient amount of examples. Corris et al. (2000) divide user comments into four categories: (i) attitudes of users and makers to dictionaries, (ii) exhaustiveness, (iii) functionality and (iv) practical considerations. The results given in sections 2.1 and 2.3 are largely reflected in the aspects arising from qualitative study categories. This is especially true when it comes to the value of examples.

2.5 User skills in dictionary use

Based on research with business students, Gromann/Schnitzer (2016) find a general lack of awareness of current specialised resources and alternative ways for processing specialised texts. Chan (2017) shows that although bilingual dictionaries determine noun countability and associated article use, learners often misinterpret dictionary information on this prop-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

erty, resulting in article errors or wrong countability judgments. Laufer (2011, p. 45) observes that "some learners could not find some target collocations in the dictionaries even though they were included there" and "sometimes learners were not aware of the fact that the collocation was unfamiliar to them and did not seek dictionary help". Chen (2017, p. 225) states:

> The participants showed inadequate dictionary use skills. They were reluctant to use the hyperlink function of the electronic dictionary to further look up relevant information or too careless to notice it, unable to distinguish between entry subsenses, inclined to choose the sense listed at the beginning of an entry, and apt to lose patience when faced with overcrowded entry information.

3. The dictionaries

3.1 Electronic dictionary of graph theory

We developed an electronic dictionary for the domain of graph theory (*GraWB* for *Graphentheoretisches Wörterbuch*). The process of lemma selection is discussed in Kruse/Heid (2021) and the microstructure in Kruse/Heid (2020). In the prototype used in the study, 452 German and 160 English lemmas have entries. The information in the microstructure depends on the headword, as each term is assigned to a semantic category such as *graph type*, *algorithm*, or *graph property*. Definitions are given for most of the German and some of the English terms. Access paths are a lemma list and a search field. Category-based access is planned, but not yet implemented. Currently, the dictionary is only accessible after login, not via a search engine.[1]

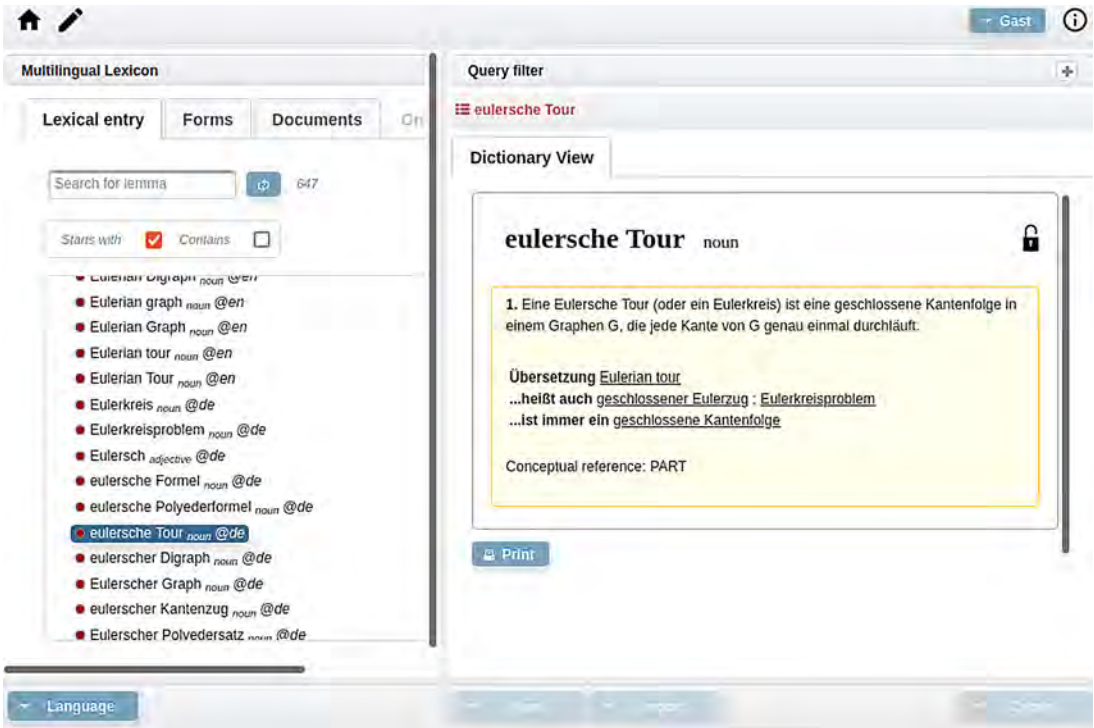


Fig. 1: Interface of GraWB

[1] https://lexo.hosting.uni-hildesheim.de/LexO/faces/loginView.xhtml.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

The dictionary is implemented in the LexO framework (Bellandi et al. 2017). LexO was created for terminologists and is optimized for desktop or tablet use. It has a panel and a tab view, however, we only use the panel view for our research. Figure 1 contains a screenshot of the user interface and an example entry. GraWB does not contain any illustrations yet. We identified a few issues during a cognitive walk-through with a usability expert, which we communicated to the developers, but which are outside the scope of this paper.

3.2 Wikipedia entries for graph theory

Wikipedia is a well-known online encyclopedia. Nevertheless, it has not been the topic of many (meta-)lexicographic analyses (Fuertes-Olivera 2009; Mederake 2014), likely because as an encyclopedia, its article structure differs massively from that of a typical (monolingual) specialised dictionary.

Although most of the articles include links to equivalents in other languages, the contents of the articles differ greatly because they are written collaboratively and independently. The links can be, however, used to detect translation equivalents. The articles consist of continuous text divided into (sub-)sections. Terms with their own article are highlighted and linked. Articles can include illustrations, videos, or sound files.

The articles can be mainly accessed via a search field and the above-mentioned links. As Wikipedia articles rank highly in most search engines, they can also be accessed this way. Inside Wikipedia, access paths like lists of articles related to a specific topic, e. g. *list of graphs* or *glossary of graph theory*, are available. As anyone can modify most parts of an article, the encyclopedia is not static but its articles may change in contents and form over time.

On a content level, the main difference between Wikipedia and GraWB is that Wikipedia does not focus on a particular domain but provides information on a vast range of domains. This may lead to cognitive overload in some users. Additional effort is needed when users have to extract the needed information from the texts.

4. Experimental setup

4.1 Study design and participants

We conducted our study with two groups of students: one in summer 2020 and the other in summer 2021. An online survey was active for around six weeks each time. The participants in both studies were given identical tasks with the exception that the first time they were only allowed to use Wikipedia as a resource (N=182), while the second time only GraWB (N=113). In the second study, the NASA task load index[2] was employed as well.

[2] https://humansystems.arc.nasa.gov/groups/TLX/.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

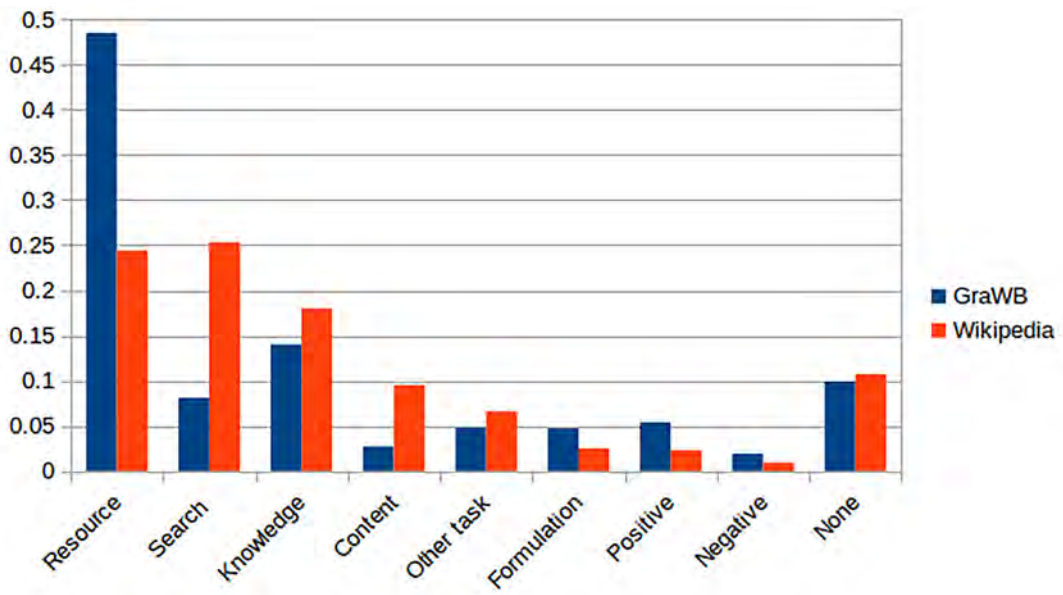


Fig. 2: Distribution of the comments over the categories

The tasks were written in German and distributed through a lecture on algebra and number theory at the University of Hildesheim attended by teacher students. The participants accessed the online survey individually. Both groups had taken similar courses in their study program and had already basic mathematical but no specific knowledge of graph theory. We also inquired about first languages, graph theory courses attended and study semester, to gain a better understanding of the impact of background knowledge. In the second survey, the participants also had to indicate whether they had already taken part in the first study which was the case for about 25% of the subjects but we do not find a significant difference (t-test) for the number of search terms, the time used and the points awarded between the two groups. We suggested taking the survey on a desktop-like surface because it allows switching between browser tabs more easily.

4.2 Tasks and types of searches

The study consists of six tasks with subtasks. They address different aspects of terminology. In task (1), the participants have to give the properties of graphs which are only represented graphically (namely a binary tree, a cycle graph and the Petersen graph), i. e. no search term is given. This allows us to evaluate whether the students describe the graphs only in general language or if they use terminology. Task (2) shows which terms the students use to describe a mathematical concept they do not know by name. Both tasks cannot be solved by mere look-up, but rather by following paths through the conceptual domain, some sort of 'ontological inference'. Tasks (3) and (4) relate to the classification of mathematical terms by Vollrath (1978): We chose a term with a homograph in general language with the same meaning as well as a term where the general language meaning differs considerably to examine how the participants apply these terms. Task (5) addresses search strategies, as we ask the participants to name algorithms for a given problem. Task (6) is to translate two sentences from English to German which together contain seven target terms. The participants obligatorily had to comment on each task. In section 5, we delve deeper into tasks (1.1), (3.1), (4) and (6.1).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

4.3 Methods for data analysis

We combine numerical scoring and an analysis of the users' textual comments to evaluate our data. Each response is given a score; the authors did all of the scoring work task-wise. In tasks (1) and (2), 1 point is given if the answer provides a graph-theoretical perspective, 0 points for a non-graph-theoretical perspective and 0.5 points for answers in between. In tasks (3) and (4), 1 point is given for a correct answer, 0 points for a wrong answer and 0.5 points for a semi-correct answer. In task (5), 1 point is given if at least two algorithms are named, 0.5 points for only one and 0 for none. In task (6), one point for each target term can be earned.

The lengths of the submitted comments range between 1 and 910 characters; the average length is 86.17 with a standard deviation of 82.19 and a median of 64. Each comment is assigned to one of the following categories:[3]

- design and content of the resource, e. g. *Article nicely manageable*
- search strategy, e. g. *I found the term by following the blue highlighting*
- prior knowledge, e. g. *I picked it up somewhere*
- mathematical content, e. g. *Looks like stochastic*
- reference to another task, e. g. *derived from the previous question*
- references to the formulation of the particular task, e. g. S*omewhat unclear to what extent the question should be answered...*
- general positive remark, e. g. *No problem with the task*
- general negative remark, e. g. *Outch, I'm so bad*
- none/other, e. g. *No comment*

Figure 2 shows the distribution of the comment categories. The majority of the comments are concerned with the resource and the search strategies used. For those, the following subcategories were found:

- no use of the resource because it would not help anyway, e. g. *Again, a translation app/website is the better option*
- lack of a helpful search term or access structure, *e. g. I found it very difficult to find a search term for this.*
- nothing found, e. g. *I somehow couldn't really find anything.*
- found a helpful article but did not understand it, e. g. *I did not understand the entries*
- description of a successful search, e. g. *The explanation was quickly found*
- general remark on the resource, e. g. *Unfortunately, no example on the page, then it would certainly be easier to understand.*
- The categories *nothing found, no access structure, not understood* mainly apply to comments by participants who did not provide a high rated answer.

5. Results and interpretation

5.1 Effectiveness: usefulness for the tasks

In terms of quantitative evaluation, we calculate the Spearman correlation between the number of searches and rewarded points (Spearman 1904). We get 0.2516 ($p<0.01$, two-tailed) for Wikipedia and 0.2136 ($p<0.05$, two-tailed) for GraWB when all sub-tasks are

[3] The example comments have been translated into English by the authors.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

counted equally, demonstrating a small correlation with either resource. The results change when we calculate the coefficient separately for task (6) and the others. With Wikipedia, we have 0.2398 ($p<0.01$, two-tailed) for tasks (1) to (5) and −0.1193 (not significant) for task (6) while it is 0.1560 (not significant) for tasks (1) to (5) and 0.2180 ($p<0.05$, two-tailed) for task (6) using GraWB. We only detect small correlations, however, the positive influence of GraWB is stronger in the translation tasks, as could be expected because it is – unlike Wikipedia – optimized for such use situations. In the following, we analyse four tasks in detail.

5.1.1 Task: describe a given graph

In task (1.1), the participants had to name properties of a binary tree, presented to them as a picture. In the Wikipedia study, 47% used a graph-theoretical description, but 75% did in the GraWB-study, while 32% took a stochastic approach with Wikipedia and only 11% with GraWB. The rest used a combination. The reason for the difference is probably that Wikipedia users got side-tracked by information about the use of graph-theoretical concepts in other domains like stochastic, while GraWB only presents graph theory.

In the following, we list the top five search terms used in that task for each of the two studies. In parentheses, we give the absolute number of participants who used this search term, followed by the average rating of the term (Likert scale, *1=very helpful*, *5=not helpful at all*) and a translation. Most terms relate to *tree*, although some are graph-theoretical and some are stochastic, as evidenced by the answers.

Wiki: *Baumdiagramm* (50, 2.42, *tree diagram*), Graphentheorie (29, 2.04, *graph theory*), Graph (18, 3, *graph*), Baum (Graphentheorie) (8, 1.75, *tree* (*graph theory*)), Eigenschaften Baumdiagramm (7, 2.57, *properties tree diagram*)

GraWB: Baum (50, 2.04), Graph (27, 3.37), binärer Baum (11, 2.18, *binary tree*), Kante (11, 2.91, *edge*), Knoten (9, 2.56, *node*)

In both trials, the majority of those who use a dictionary provides a graph-theoretical perspective (63% for Wikipedia and 80% for GraWB) with GraWB reaching a higher percentage, as expected. For those who did not use a dictionary we find a wider disparity although we had expected similar results: In the Wikipedia study 66% of those not using a dictionary got 0 points but only 28% without GraWB did. A reason could be the small number of participants not using a dictionary (65 for Wiki, 32 for GraWB) which makes it difficult to draw further conclusions.

5.1.2 Task: apply a term – part 1

In task (3), the participants had to answer the question *What is N adjacent to?* for an indicated node *N* in a labeled Petersen graph. The German formulation of the task (*Wozu ist N adjazent?*) contains the interrogative adverb *wozu*, due to the syntactic construction of the adjective (*x ist adjazent zu y*). This syntactic information is not derivable from the respective Wikipedia articles. The answers of three participants in each study indicate that they misunderstood *wozu* as *for which purpose*, which may be seen as evidence for the need to give constructional information also in specialized dictionaries. In the following, we present again the search terms and values:

Wiki: adjazent (127, 1.35), Nachbarschaft (Graphentheorie) (19, 1.05), Adjazenzmatrix (10, 2.1), adjazent⟶Nachbarschaft (Graphentheorie) (6, 1.5), Pentagramm (3, 3), Nachbarschaft (3, 2.67)

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

GraWB: adjazent (100, 1.15), adjazente Ecke (20, 1.25), adjazenter Knoten (14, 1.29), benachbart (5, 2), Ecke (4, 1.25)

We assume that those students who give *Nachbarschaft* (*Graphentheorie*) used *adjazent* as a search term because Wikipedia automatically forwards *adjazent* to *Nachbarschaft* (*Graphentheorie*). In the Wikipedia study, 98.35% used the resource, while 90% did in the GraWB study. In both studies, those who did not use a dictionary solved the task correctly with prior knowledge, except for two GraWB-users who claim that they could not find anything helpful. We see that the resources support finding the meaning of terms only used in terminology with Wikipedia having a slight advantage as an already established resource which students are familiar with.

5.1.3 Task: apply a term – part 2

In task (4), the participants answered the question *What are the leaves of the graph?* for a given labeled graph. *Leaf* is an example of a mathematical term that is homonymous with a general language word, but its meaning cannot be easily derived.

Wiki: Blätter und innere Knoten in der Graphentheorie (32, 1.25), Baum (Graphentheorie) (20, 2.2), Blätter Graph (17, 2.29), Blätter Graphentheorie (17, 2.53), Blätter Graphen (16, 3.31)

GraWB: Blatt (87, 1.60), Blatt eines Baumes (20, 4.15), Grad der Ecke (15, 1.33), Blätter (8, 4), Baum (8, 1.63)

When searching Wikipedia for *Blätter Graph* or *Blätter Graphen* the article *Baum* (*Graphentheorie*) is suggested while most search engines return the Wikipedia article *Blätter und innere Knoten in der Graphentheorie* as a prominent result. The first two options in GraWB when typing *Bl* are *Blatt* and *Blatt eines Baumes.* Unfortunately, the GraWB-search does not include lemmatization, therefore *Blä* yields no results. In the Wikipedia study, 94% used the dictionary and 84% did in the GraWB study.

5.1.4 Task: translate a sentence

The participants were asked to translate the sentence *A square grid graph may have a spanning tree* to German with the target terms *square grid graph* and *spanning tree.*

Wiki: spanning tree (41, 1.98), square grid graph (20, 2.7), grid (15, 2.53), square grid (13, 2.85), A square grid graph may have a spanning tree (12, 4.12)

GraWB: spanning tree (58, 1.57), square grid graph (57, 1.47), Gittergraph (9, 2.22), Spannbaum (6, 3), square grid (4, 4).

As expected, the target terms are the most popular search terms. When searching German Wikipedia for *spanning tree*, its equivalent *Spannbaum* appears among the top five suggestions. Analogously, *square grid graph* leads to *Gittergraph. S*ome participants searched for the complete sentence (cf. translation engine metaphor) but were not very satisfied with the results. The participants are slightly more satisfied with the outcomes in GraWB, but this is probably because Wikipedia is not optimized for translation. Only 52% used Wikipedia in this task, while 74% used GraWB.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

5.2 Design aspects of the dictionary

In this section, we present further aspects mentioned by the participants. One example, already introduced in section 4.3, is the access path, as several participants struggle to come up with a suitable search term. Different reasons cause this problem: Some users have no notion of how to complete the task and others have a concept in mind but do not know the corresponding terminology. A graphical access structure with pictures could help both. At least in graph theory, such approaches appear feasible, while it remains to be investigated if it also applies to other subdomains.

Another idea is to include non-terminological quasi-synonyms of terms in the lemma list and introduce links to the relevant terms. For example, the term *isomorphic* denotes an identity of graphs even if they are visualized differently. An entry *same as*, appropriately marked as a general language item, could link to an entry for *isomorphic*, relating an everyday non-terminological expression to the actual term. It has to be evaluated from a didactic perspective if such a device aids or hinders acquiring the right terminology as it can be problematic for those who acquire new terms as opposed to those who recall terms they have already learned.

The users comment differently on links: Some regard them as helpful while others are annoyed by clicking on several links until they reach their goal. Alternative access structures may help to reduce the number of links that must be used, e. g. clusters like those introduced by *Wortgeschichte digital*.[4]

As graph theory strongly relies on visualizations, it is natural that the users would have liked the dictionaries to include (more) illustrations. While no positive effect of illustrations could be found for learning languages (cf. section 2.2) it could be explored in further research if that also holds for terminology.

We further see preferences for using similar strategies as in search engines while other participants indicate that they developed a routine of dictionary use in the course of the study. Sentiment analysis of user feedback shows that the overall satisfaction is highly task-dependent. In tasks (1), (2) and (5), the comments are slightly more positive in the Wikipedia study, while in the others, GraWB has more positive comments.

6. Limitations of the study

Due to the pandemic, we could not use tracking software but had to rely on voluntary disclosure by the users. Thus, we cannot be sure if they visited other articles but did not give them in the form they had to fill, as some comments mention translation tools for task (6) in the Wikipedia study. Another caveat is that some users gave their actual search terms, while others gave the first article they were forwarded to or clicked on. Further, some comments indicate that the users access Wikipedia articles through a search engine. While within Wikipedia, results only depend on the search term, most search engines personalize the ranking of the results and thus we can only vaguely retrace the search paths. Furthermore, some search results are highlighted by the engines with a picture, mostly a Wikipedia article, so that users are tempted to click on it without investigating further results. Finally, we

[4] https://www.zdl.org/wb/wortgeschichten.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

did not vary the order of the tasks and therefore carry-over effects may appear. For replication, it has to be considered that Wikipedia articles change and that all participants may access different versions.

7. Conclusion and future work

The present study can be the starting point of a design-based research study on the development of terminological resources in mathematics. That way the ideas given by the students can be further investigated. The same applies to other access paths where articles are suggested after an associated search term. Further ideas are to develop the mixed-methods approach in either a purely quantitative or a purely qualitative study. Not much work has so far gone into the user-centered design of specialized dictionaries. The present study, although carried out on a particular subdomain of mathematics, led to several conclusions that may be valid more generally.

The comparison between our dictionary of graph theory and Wikipedia showed that a specialized dictionary avoids that users get side-tracked by information on related domains; and that the particularized microstructure of a dictionary often allows for more efficient access to individual items than the explanatory texts of the encyclopedia and the different types of links contained therein. We also learned that both types of resources require access by lemmas or related items and that students who do not know which search term to use have difficulties in accessing the necessary information. The encyclopedia articles tend to be somewhat more helpful in such situations, but are still far from satisfactory. Whether the inclusion of lay terminology into the dictionary, accompanied by references to specialized terms of the domain, is an appropriate solution remains to be investigated. Likewise, more research is necessary to understand whether some kind of picture-based search is feasible for a limited and visually rather straightforward domain like graph theory, to cater for such situations.

While the translation advantages of the dictionary are expected, the fact that some users deal with the dictionary the same way as with an online translation system (enter full sentences, wait for the translation result) was less expected, yet it shows the extent to which generic online tools influence user expectations. Future specialized e-dictionaries will indeed have to offer specific advantages (e. g. accuracy, ease of access, classification, functions of an ontological categorization, equivalents) over generic online tools to be recognized and used by the public.

In terms of methodology, we see a need for refined methods for user-centered analyses of specialized dictionaries: In a quantitative study, further data could be collected, e. g. not based upon voluntary disclosure but on log files and by measuring the exact time for each task or search. A qualitative study could include extensive interviews with the participants to gain further insights into their working procedures. These would not only be interesting from a lexicographic but also from a didactic perspective. Another approach could be to let the tasks be done twice by the same group as a pre- and post-study. It might be also interesting to introduce the dictionary to the participants but let them choose if they use Wikipedia, our resource, or any other device.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

References

Aust, R./Kelley, M. J./Roby, W. (1993): The use of hyper-reference and conventional dictionaries. In: Educational Technology Research and Development 41 (4), pp. 63–74. http://doi.org/10.1007/BF02297512.

Bellandi, A. et al. (2017): Developing LexO: a collaborative editor of multilingual lexica and termino-ontological resources in the humanities. In: Proceedings of Language, Ontology, Terminology and Knowledge Structures Workshop (LOTKS 2017). Montpellier, France. https://aclanthology.org/W17-7010.

Boonmoh, A. (2021): Use of dictionaries and online tools for reading by Thai EFL Learners in a naturalistic setting. In: Lexikos 31 (1). http://doi.org/10.5788/31-1-1645.

Chan, A. Y. W. (2014): Using LDOCE5 and COBUILD6 for meaning determination and sentence construction: What do learners prefer? In: International Journal of Lexicography 27 (1), pp. 25–53. http://doi.org/10.1093/ijl/ect034.

Chan, A. Y. W. (2017): The effectiveness of using a bilingualized dictionary for determining noun countability and article selection. In: Lexikos 27 (1). http://doi.org/10.5788/27-1-1399.

Chen, Y. (2017): Dictionary use for collocation production and retention: a CALL-based study. In: International Journal of Lexicography 30 (2), pp. 225–251. http://doi.org/10.1093/ijl/ecw005.

Corris, M. et al. (2000): Bilingual dictionaries for Australian languages: user studies on the place of paper and electronic dictionaries. In: Heid, U. et al. (eds.): Proceedings of the 9th EURALEX International Congress. Stuttgart, pp. 169–181.

Dziemianko, A. (2012): On the use(fulness) of paper and electronic dictionaries. In: Granger, S./Paquot, M. (eds.): Electronic lexicography, pp. 319–342. http://doi.org/10.1093/acprof:oso/9780199654864.003.0015.

Dziemianko, A. (2015): Colours in online dictionaries: a case of functional labels. In: International Journal of Lexicography 28 (1), pp. 27–61. http://doi.org/10.1093/ijl/ecu028.

Farina, D. M. T. C./Vrbinc, M./Vrbinc, A. (2019): Problems in online dictionary use for advanced Slovenian learners of English. In: International Journal of Lexicography 32 (4), pp. 458–479. http://doi.org/10.1093/ijl/ecz017.

Fuertes-Olivera, P. A. (2009): The function theory of lexicography and electronic dictionaries: Wiktionary as a prototype of collective free multiple-language internet dictionary. In: Bergenholtz, H./Nielsen, S./Tarp, S. (eds.): Lexicography at a crossroads: dictionaries and encyclopedias today, lexicographical tools today. (= Linguistic insights, 1424-8689 v. 90). Bern/New York, pp. 99–134.

Fuertes-Olivera, P. A. (2016): A cambrian explosion in lexicography: Some reflections for designing and constructing specialised online dictionaries 1. In: International Journal of Lexicography 29 (2), pp. 226–247. http://doi.org/10.1093/ijl/ecv037.

Gromann, D./Schnitzer, J. (2016): Where do business students turn for help? An empirical study on dictionary use in foreign-language learning. In: International Journal of Lexicography 29 (1), pp. 55–99. http://doi.org/10.1093/ijl/ecv027.

He, Y. (2017): Dictionary user research in the digital era: value of user-generated content. In: Proceedings of ASIALEX 2017, pp. 676–693. https://asialex.org/pdf/Asialex-Proceedings-2017.pdf.

Heid, U./Zimmermann, J. T. (2012): Usability testing as a tool for e-dictionary design: collocations as a case in point. In: Fjeld, R. V./Torjusen, J. M. (eds.): Proceedings of the 15th EURALEX International Congress. Oslo, pp. 661–671.

Hult, A.-K. (2014): The authentic voices of dictionary users – viewing comments on an online learner's dictionary before and after revision. In: Abel, A./Vettori, C./Ralli, N. (eds.): Proceedings of the 16th EURALEX International Congress. Bolzano, pp. 237–247.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Kemmer, K. (2014): Rezeption der Illustration, jedoch Vernachlässigung der Paraphrase? Ergebnisse einer Benutzerbefragung und Blickbewegungsstudie. In: Müller-Spitzer, C. (ed.): Using online dictionaries, pp. 251–278. http://doi.org/10.1515/9783110341287.251.

Khairiah, D. (2021): Definition model for plant and animal AME lemmas in KBBI V: a user study. In: Proceedings of ASIALEX 2021, pp. 142–149. https://asialex.org/pdf/Asialex-Proceedings-2021.pdf.

Kruse, T./Heid, U. (2020): Lemma selection and microstructure: definitions and semantic relations of a domain-specific e-dictionary of the mathematical domain of graph theory. In: Gavriilidou, Zoe/Mitsiaki, Maria/Fliatouras, Asimakis (eds.): Lexicography for Inclusion: Proceedings of the 19th EURALEX International Congress, 7–9 September 2021, Alexandroupolis. Volume 1. Alexandroupolis, pp. 227–233.

Kruse, T./Heid, U. (2021): From term extraction to lemma selection for an electronic LSP-dictionary in the field of mathematics. In: Electronic lexicography in the 21st century: post-editing lexicography, pp. 572–587.

Laufer, B. (2011): The contribution of dictionary use to the production and retention of collocations in a second language. In: International Journal of Lexicography 24 (1), pp. 29–49. http://doi.org/10.1093/ijl/ecq039.

Laufer, B./Hill, M. (2000): What lexical information do L2 learners select in a CALL dictionary and how does it affect word retention? In: Language Learning & Technology 3 (2), pp. 58–76.

Laufer, B./Levitzky-Aviad, T. (2006): Examining the effectiveness of 'Bilingual Dictionary Plus' – a dictionary for production in a foreign language. In: International Journal of Lexicography 19 (2), pp. 135–155. http://doi.org/10.1093/ijl/eck006.

Lew, R. et al. (2018): Competition of definition and pictorial illustration for dictionary users' attention: An Eye-Tracking Study. In: International Journal of Lexicography 31 (1), pp. 53–77. http://doi.org/10.1093/ijl/ecx002.

Li, L./Xu, H. (2015): Using an online dictionary for identifying the meanings of verb phrases by Chinese EFL learners. In: Lexikos 25. http://doi.org/10.5788/25-1-1295.

Lorentzen, H./Theilgaard, L. (2012): Online dictionaries – how do users find them and what do they do once they have? In: Fjeld, R. V./Torjusen, J. M. (eds.): Proceedings of the 15th EURALEX International Congress. Oslo, pp. 654–660.

Mederake, N. (2014): Artikel der Wikipedia aus lexikografischer und textlinguistischer Perspektive. In: Mann, M. (ed.): Digitale Lexikographie: ein- und mehrsprachige elektronische Wörterbücher mit Deutsch: aktuelle Entwicklungen und Analysen. (= Germanistische Linguistik 223–224). Hildesheim, pp. 229–249.

Miller, J./Kwary, D. A./Setiawan, A. W. (2017): Koalas, Kiwis and Kangaroos: the challenges of creating an online Australian Cultural Dictionary for learners of English as an additional language. In: Lexikos 27 (1). http://doi.org/10.5788/27-1-1405.

Nesi, H. (2015): Thirty years of user studies – and what we still need to find out. In: Li, L./Mckeown, J./Liu, L. (eds.): Proceedings of ASIALEX 2015 Hong Kong: words, dictionaries and corpora: innovations in reference science, pp. 1–8.

Spearman, C. E. (1904): The proof and measurement of association between two things. In: The American Journal of Psychology 15 (1), pp. 72–101. http://doi.org/10.2307/1412159.

Tarp, S. (2008): Lexicography in the borderland between knowledge and non-knowledge. General lexicographical theory with particular focus on learner's lexicography. Tübingen.

Töpel, A. (2014): Review of research into the use of electronic dictionaries. In: Müller-Spitzer, C. (ed.): Using online dictionaries, pp. 13–54. http://doi.org/10.1515/9783110341287.13.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Vollrath, H.-J. (1978): Lernschwierigkeiten, die sich aus dem umgangsprachlichen Verständnis geometrischer Begriffe ergeben. In: Bauersfeld, H./Otte, M./Steiner, H.-G. (eds.): Lernschwierigkeiten im Mathematikunterricht. Bielefeld, pp. 57–73.

Contact information

Theresa Kruse
University of Hildesheim
kruset@uni-hildesheim.de

Ulrich Heid
University of Hildesheim
heidul@uni-hildesheim.de

Acknowledgments

We would like to thank Andrea Bellandi, Simone Marchi and Silvia Piccini from the University of Pisa for their support by setting up LexO, as well as Jürgen Sander and Martin Kreh for their support in executing the study and Lea Wöbbekind for her insights on usability.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Silga Sviķe

SURVEY ANALYSIS OF DICTIONARY-USING SKILLS AND HABITS AMONG TRANSLATION STUDENTS

Abstract The paper presents the results of empirical research conducted with students from the Faculty of Translation studies of Ventspils University of Applied Sciences (VUAS) in Latvia. The study investigates the habits and practices concerning the use of dictionaries on the part of translation students, as well as types of dictionaries used, frequency of use, etc. The study also presents an insight into the evaluation of the usefulness of dictionaries by Latvian students. The research describes the advantages and disadvantages of dictionaries used by the respondents, the importance of the preface and the explanation of the terms and abbreviations used in dictionaries. The research conducted, as well as the insights, results and recommendations presented, will be relevant for the lexicographic community, as it reflects the experience of one Latvian University to improve the teaching of dictionary use and lexicographic culture in this country and to complement dictionary use research with the Latvian experience.

Keywords Dictionaries; dictionary use; translators; translation studies; translation tools; survey

1. Introduction

An essential aspect of the translation process is the search for information in various resources, such as dictionaries, specialised literature and parallel texts. Checking word meanings in a dictionary is a fundamental stage of translation, especially when translating from or into a second foreign language, rather than translating from or into a first foreign language, as translators may not feel as confident in their use of a second language. Reasons for using a dictionary may vary, for instance, to ascertain the spelling of a word, to obtain information on the grammatical category of a word, or to obtain an explanation or translation. It is important to be aware of the importance of this process since dictionaries provide translators with valuable additional information, such as different meanings, special usages, etc. Atkins/Varantola (1998, p. 98) divide this information search into 2 types: primary information (when the first look-up in a search might be a quest for a second language (L2) translation) and secondary information (once the L2 word has been found, the search may involve looking to confirm the chosen translation or to obtain grammatical or other additional information about it. Translation students are expected to be particularly good at dictionary-using skills and to be knowledgeable about different dictionaries, their types, their structure, their functions and their use in the translation process. As all paper dictionaries have not yet been digitised, this applies to printed dictionaries as well as electronic ones.

This study describes the experience of using dictionaries in translation studies, as well as the dictionary-using habits of translation students, their frequency of use, and problems or difficulties faced in the process of dictionary use, especially when learning a foreign language and carrying out translation tasks involving a first or a second foreign language. The study aims to provide an insight into the author's observations and experiences on how translation students learn to use dictionaries, use them in practical assignments, what their dictionary usage habits are and how these observations can further improve the quality of

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

dictionaries in Latvian, according to the study's findings. The objective of the study is based on the assessment of the skills of dictionary users and the usefulness of dictionaries.

The main research methods applied in the study are a questionnaire, which is the most common approach for collecting data on dictionary use (Nesi 2013), the analysis of collected data and a descriptive method, based on the empirical study involving the observation of translation students during translation tasks, the analysis of feedback reported by students after practical tasks, as well as the analysis of their tasks, completed during study courses such as *Translation* and *Terminology and Lexicography*.

2. The need for researching dictionaries and dictionary-using habits in translation studies

The ability to search for information in various support tools, including different types of dictionaries, is becoming increasingly important in the everyday life of translators as their work routine relies significantly on modern technology. Translators must be able to evaluate all the translation support tools, to select the information needed for each specific case. Research of translation tools and dictionaries (especially to assess their usefulness) should be at the heart of planning and designing any new lexicographic sources and should contribute significantly to the publishing of new lexicographic sources (Dringó-Horváth 2014, pp. 218 f.; Hartmann 2000, p. 390). Already in 1987 R. R. K. Hartmann (1987, p. 11) proposed several reasons why dictionaries should be researched, given that reviews mention excessively diverse requirements for dictionaries while it often remains unclear to dictionary users what kind of information to expect. There are also uncertainties among language teachers about what dictionaries to recommend to their students for the language learning process. Dictionary research is also particularly important because the form, type and content of modern dictionaries are currently evolving and rapidly changing: paper dictionaries are still being digitised, dictionaries are being developed with new content, and the content of the information available in electronic dictionaries is changing as well, as they are improved, supplemented, etc. In addition, skills related to the use of dictionaries are already acquired independently or they need to be regularly improved and supplemented in schools; there should be a constant interest in the range of new dictionary editions, electronic dictionary updates and improvements in order to select at a metacognitive level the necessary edition, one which can be of assistance in translation studies or other instances of dictionary use. In particular, the results of studies on the use of dictionaries help find out what kind of lexicographic publications will be needed in the future. This is especially important for dictionaries in Latvian, as in practice they are still relatively seldom reviewed. Such user feedback is important to developers of these resources, as well as language and special course teachers. Moreover, Latvian students' dictionary use habits and skills are also very rare as a research topic. At present, more and more attention is being paid to this topic, not only within educational or research institutions but also between them. There are specific conferences held specifically dedicated to dictionary didactics issues, such as the conference "New Challenges in Dictionary Teaching" (organised by Università degli Studi Roma Tre in October 2021), as well as papers presented at conferences dealing with specific translation issues, such as "Meaning in Translation: Illusion of Precision" (organised by Riga Technical University in May 2018).

There are several available studies on the use of dictionaries during the process of learning a foreign language, in which the primary research group are secondary school students

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

(Dringó-Horváth 2011, 2012, 2014; Hessky 2009; Świątkiewicz-Siklucka 2008). There also exist several studies on how university students, including translators, use dictionaries and other translation tools (e.g., Atkins/Varantola 1998; Hamouda 2013; Hartmann 1987, 2000; Kodura 2016; Paradowska 2020; Tono 2012). However, there is a lack of such data concerning Latvian university students. Additionally, dictionaries containing Latvian as one of the working languages have been less studied in relation to this issue.

Research on dictionary-using strategies in the process of second foreign language acquisition in Latvian secondary schools has shown that the ability to use dictionaries during foreign language lessons has proved to be very useful, as it was concluded that acquired skills apply not only to the learning of a foreign language, but these skills can be of significant help during further studies. In addition, it was concluded that students must first acquire the skill to work with printed dictionaries, which is the very basis for further use of electronic dictionaries. Printed dictionaries should be available in the classroom to make it a habit to use them. In their absence, students should be allowed to use electronic dictionaries. Dictionary-using skills will certainly strengthen self-directed language learning, as those students who have mastered this strategy will be able to continue learning the language outside the classroom and will know how to use a dictionary to support language learning and use. The data of the survey, in which 42 Latvian foreign language teachers participated, shows that it is necessary to start working with a dictionary for learning a foreign language in the 4th grade, and 61.9% of the survey respondents (foreign language teachers in Latvian schools) indicated that they had included the topic *Learning strategy – working with a dictionary* in the curriculum. However, a third of the respondents (31%) had not yet implemented or planned such a topic in the curriculum. As a result of the research, a recommendation has been made to include the work with dictionaries as a language learning strategy in the curricular program of foreign language teachers in Latvian universities, so that teachers can subsequently then transfer this knowledge and skills to students during foreign language lessons. (Sviķe in prep.)

Dictionary-using skills are also mentioned in various documents listing skills and knowledge required for translators. For example, competencies needed for the professional activity of a translator are specified in Paragraph 3 of the Standard for the Profession of Translator. Among other things, Paragraph 3(5) of the Standard specifies the ability to select and evaluate lexicographical resources in order to perform a high-quality translation (The Standard for the Profession of Translator 2012). In The EMT Competence Framework 2017, the European Master's in Translation (EMT) network Board defines five main areas of competence that translation program students need to be taught in universities: language and culture competence, translation competence, technological competence, personal and interpersonal competence, service-providing competence. These competences should be considered complementary and equally important in providing the translation service, which is the main goal of the translation process. (Taudic/Krause 2017) Dictionary-using skills would fall within the translation competence, which describes that students are able to select and critically evaluate resources for translation, choose appropriate translation strategies, and use different translation tools and techniques in translation. (ibid., pp. 7 f.). For students to be able to use these reference materials (i.e., dictionaries) effectively, it is first necessary to get an understanding of how they use these resources in the process of translation and what is their user profile. In this context, the present study was carried out to find out significant data regarding the dictionary-using skills and dictionary usage habits on the part of translation students.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

3. Structure of the empirical research

This chapter provides an insight into the structure of the research on dictionary use conducted to elucidate dictionary-using skills of translation students by analysing the data provided by the students of the Faculty of Translation Studies (FTS) of the Latvian regional university Ventspils University of Applied Sciences (VUAS). The purpose of this task was to create a profile of the VUAS FTS students as dictionary users, to find out their dictionary-using habits and summarise the main difficulties they encounter when using dictionaries as well as provide possible solutions. The research results could be a good basis for dictionary compilers – lexicographers – to improve dictionaries with Latvian as one of the languages on the one hand, and translation students' educators on the other hand.

To implement this idea, the following tasks were set: to obtain and compile information on translation students' dictionary-using skills (frequency of use, difficulties in the search process, possible causes of these difficulties, the first opportunity to obtain dictionary-using skills, etc.), taking into consideration dictionaries commonly used in language lessons, their benefits, and suggestions on how to improve existing dictionaries.

3.1 Characteristics of the research participants and insight into the content of translation studies

This chapter describes the answers to the survey given to students from the first to the fourth year of VUAS FTS and their evaluation of dictionaries. A questionnaire was used to obtain the research data. A total of 78 translation study students (70% of all VUAS FTS students) responded to the questionnaire. Second-year students need to acquire several courses that include learning different translation strategies and using dictionaries, both printed and electronic (e. g., the course *Terminology and Lexicography*), as well as carry out practical activities using other translation tools (such as CAT tools – *Memsource*, *SDL Trados Studio*) in the course *Computer-Aided Translation*. Students use different types of dictionaries from the very beginning of their studies when they have mastered both their mother tongue and the first and second foreign languages, as well as completed other theoretical courses. Several practical courses on the use of dictionaries, which evidently require the use of dictionaries and other tools, are also run by the author of this article – such as the practical translation courses *Translation of contracts*, *Translation of documents*, and *Translation of technical texts*. In the following study, there are also some observations about what the author of the present article has observed and summarised during the courses that require work with dictionaries.

3.2 Methodology of the research

The research questionnaire consists of 28 questions adapted from the R. R. K. Harmann (1999) questionnaire. Students' responses were obtained via survey forms, which were filled in electronically in the spring semester of the 2021/2022 academic year. The responses were analysed using data analysis functions exported to Excel spreadsheets. The content of the questionnaire is as follows:

- Academic information about the respondents
- Types of dictionaries most frequently used

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

- Reasons for using a dictionary
- Use and knowledge of usage guides
- Frequency, aim and reason of dictionary use
- Main problems and causes of difficulties when searching for information in dictionaries, as well as possible solutions
- Dictionaries recommended by respondents as useful for translation studies

The data obtained was collected electronically and a quantitative and qualitative analysis (described below) was performed. At the end of the research, suggestions for the improvement of existing dictionaries with Latvian are provided. Conclusions drawn during the analysis of the results are also summarised.

3.3 Analysis of the research data

The data obtained during the research provides important information on the dictionary-using habits of translation students, essential both in the development of the content of the first and second foreign language courses, as well as in the evaluation of dictionaries (identifying difficulties, offering solutions), supplementing and improving their content in the light of the results of the data analysis.

Before analysing the answers to the specific questions, basic information about the respondents was summarised. Out of the total number of respondents (78), 24.4% are 1st-year VUAS FTS students, 28.2% are 2nd-year students, 25.6% are 3rd-year students, and 21.8% are 4th-year students. 1st-year respondents might be relatively described as beginners in translation studies, while the rest of the students can be characterized as having an intermediate level. This division is based on the fact that starting from the second year of their studies, students begin to take specialised study courses.

3.3.1 Where and how have respondents learnt to work with dictionaries

Out of the total number of respondents, 57.74% answered that they have mastered the work with dictionaries in VUAS, 21.08% that they have acquired the skill independently, while 14.1% have mastered the work with dictionaries at school and 4% of respondents have admitted to not having yet acquired such skills. All respondents who indicated the answer "I have not mastered the skills to work with a dictionary" are 1st-year students. These results are a clear signal to the creators of study programmes and lecturers of translation studies. As the data shows, the assumption that students arrive at the university with good knowledge in the field of dictionary use, know their types and how to use them, and have the skills to work with them because they learnt it at school, is misleading and erroneous. Although "the ability to translate a word using different digital dictionaries, compare translations and choose the best option" (VISC 2020, p. 27) is mentioned as one of the most important skills in secondary education, it should be noted that this may not be the case, as the survey results prove.

3.3.2 Self-assessment of dictionary use

Respondents' self-assessment of their ability to work with dictionaries according to a scale from 1 (the lowest) to 5 (the highest) is as follows: 2.6% of respondents assess their skills in

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

working with dictionaries with "2", the same percentage of respondents (42.3%) assess their ability to work with dictionaries with ratings "3" and "4", and only 12.8% of respondents believe they know how to work with dictionaries exceptionally well – "5". The data according to the division of students by levels – beginners (1st year) and intermediate level students (2nd to 4th year) – is shown in Table 1. The self-assessment data of the respondents summarised in the table shows that the level of mastery in using dictionaries is higher in senior courses. It is assessed by the respondents themselves and could be explained by the fact that these skills are indeed developed and improved during the studies, and it is only logical that more frequent work with lexicographic resources would enrich and improve their usage skills.

Self-assessment "How can I work with a dictionary?" (1–5)	**1st-year respondents**	**2nd to 4th-year respondents**
1	0%	0%
2	10%	1,7%
3	55%	39%
4	30%	44,1%
5	5%	15,3%

Table 1: Students' self-assessment of their dictionary-using skills

3.3.3 Acquaintance with the preface and users' guide sections of dictionaries

Out of all respondents, 38.5% answered the question "Do you read the preface or users' guide before using the dictionary?" with an affirmative reply, while the majority of respondents (64.1%) chose the answer "No, I do not read the information mentioned in the introduction to the dictionary.". In turn, 2.6% of respondents chose the answer "Other", indicating in the comments that in case of necessity they study the users' guide when using the dictionary if the search principle is not clear. Unfortunately, these percentages are unsatisfactory, as they show once again that, due to a variety of reasons, respondents are reluctant or not interested in reading the users' guide for a particular dictionary, although at the end of the questionnaire they acknowledge that many dictionaries are not easy for users to understand. This would probably not have been the case if students had previously carefully studied the users' guide. An overwhelming majority of 1st-year students (90%) answered that they never read a dictionary's preface or users' guide, while only 55% of respondents in the 2nd to 4th-year group answered "No, I do not read this information". This result certainly indicates that dictionary compilers need to think about how to make this section of the dictionary more "attractive" to users. It should be noted here that several of the bilingual dictionaries available on *Letonika* (https://www.letonika.lv/) do not even have users' prefaces containing explanatory notes and a list of abbreviations employed. Therefore, the skills acquired with printed dictionaries to read the meaning of the explanations they contain to understand the content are the basis for the use of electronic ones. The responsibility for introducing the use of dictionaries, of course, also lies with the lecturers of lexicography and practical translation courses.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

3.3.4 Types of dictionaries used, frequency of use and reason for using dictionaries

There were questions about the specific types of printed or electronic dictionaries students use in translation studies (see Table 2). The most commonly used type of dictionary is a bilingual or multilingual dictionary of different languages (87%). In similar studies, bilingual translation dictionaries rank first in terms of usage (Atkins/Varantola 1998, Lew 2004). The next most frequently used dictionary among the respondents is the standard monolingual Latvian dictionary (80%), followed by English-language monolingual dictionaries used by slightly more than half of the respondents, and finally German and Russian monolingual dictionaries (as German and Russian are the second foreign languages for VUAS FTS students). Among other useful literature and tools, dictionaries of various other languages (e. g., Spanish, French or Japanese) were mentioned.

Most used types of dictionaries	Respondents' answers (%)
Bilingual or multilingual dictionaries	87,2%
Latvian monolingual dictionaries	80%
English monolingual dictionaries	55%
German monolingual dictionaries	30%
Russian monolingual dictionaries	5%
Other	2,6%

Table 2: Types of dictionaries most used by FTS students

Table 3 (see below) describes the frequency of use of different types of dictionaries. The data presented in the table shows that intermediate level respondents use different types of vocabulary more often, which can be explained by the fact that they study practical translation courses where it is necessary to use these resources. In the groups of Russian and German dictionaries, the high percentage of infrequent use is made up of students for whom these languages are not the second foreign language (working language).

Frequency of use of dictionaries	1st-year course (%)	2nd to 4th-year course (%)
Bilingual or multilingual dictionaries	Every day – 26,3%; several times a day – 15,8%; once a week – 21,1%; less than once a week – 36,8%	Every day – 34,6%; several times a day – 26,9%; once a week – 23,1%; less than once a week – 17,9%
Standard Latvian monolingual dictionary	Every day – 15,8%; several times a day – 36,8%; once a week – 21,1%; less than once a week – 26,3%	Every day – 23,1%; several times a day – 35,9%; once a week – 24,4%; less than once a week – 17,9%
English monolingual dictionary	Every day – 15,8%; several times a day – 26,3%; once a week – 26,3%; less than once a week – 31,6%	Every day – 24,4%; several times a day – 30,8%; once a week – 23,1%; less than once a week – 21,8%
German monolingual dictionary	Every day – 1,9%; several times a day – 7,1%; once a week – 7,1%; less than once a week – 83,9%	Every day – 9,8%; several times a day – 16,4%; once a week – 9,8%; less than once a week – 63,9%

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Frequency of use of dictionaries	1st-year course (%)	2nd to 4th-year course (%)
Russian monolingual dictionary	Every day – 12,5%; several times a day – 0%; once a week – 12,5%; less than once a week – 75%	Every day – 4,2%; several times a day – 5,6%; once a week – 19,7%; less than once a week – 71,8%

Table 3: Frequency of use of dictionaries

Respondents mostly use bilingual or multilingual dictionaries for translation into a foreign language, while the standard Latvian monolingual dictionary is used for translations into the mother tongue (39.5%). It should be noted that the vast majority of respondents have Latvian as their mother tongue. A relatively similar percentage results for the use of relevant monolingual dictionaries (in English, German, and Russian, which are the languages of specialisation), most commonly used for translating into foreign languages, carrying out written exercises, and reading comprehension tasks. Respondents use all the dictionaries mentioned in the survey much less often for carrying out such tasks as listening and speaking comprehension. The results show that the standard Latvian monolingual dictionary is not used at all when dealing with these types of tasks. This can be explained by the fact that Latvian is the mother language of the majority of respondents. In the largest electronic dictionaries in Latvian or in dictionary sites such as *Tēzaurs* (https://tezaurs.lv/) and *Mūsdienu latviešu valodas vārdnīca* (https://mlvv.tezaurs.lv/) audio recordings or pronunciation are not given. It is possible to listen to the pronunciation of Latvian words in the electronic dictionary *e-PUPA* (http://epupa.valoda.lv/), but the number of words included in this dictionary is very small, so its support in this issue is limited. Adding audio recordings of words would be one of the most pressing tasks for the authors of modern Latvian dictionaries (both general and specialised), as there is a lack of such functionality in electronic Latvian dictionaries.

Aim of use	Bilingual/ multilingual dictionaries (%)	Latvian monolingual dictionary (%)	English monolingual dictionary (%)	German monolingual dictionary (%)	Russian monolingual dictionary (%)
Writing tasks	24,4%	20,9%	18,95%	29%	25,6%
Reading comprehension	8,2%	25,2%	20,05%	21,6%	23,7%
Listening tasks	2,1%	1,9%	7,4%	1%	3,4%
Speaking comprehension	1,2%	-	3,3%	1,8%	0,6%
Translation into the mother tongue	22,45%	39,5%	13,8%	10,9%	21%
Translation into the foreign language	41,65%	12,5%	36,5%	35,7%	25,7%

Table 4: Aim of dictionary use

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

The purpose for which bilingual or multilingual dictionaries are used in translation studies is mainly to search for unknown words (92.3%), but a very small number of respondents use these dictionaries to find the pronunciation of words (4.2%) and examples of their use (3.5%). All the monolingual dictionaries mentioned in the questionnaire are mostly used for finding definitions and explanations. Concerning the standard Latvian monolingual dictionary, it has been found that 89.7% of respondents use it to look up definitions and explanations, 73.1% use it to check the spelling of words, while 46.2% look for examples of use. 93.6% of respondents use the English monolingual dictionary to find definitions, explanations, and spelling (64.1%), while 50% of respondents look for examples of use. 78% of respondents use the German monolingual dictionary to search for definitions and explanations of words, 45.8% of respondents use it to check spellings, 42.2% look for examples, and 40.7% of respondents to search for grammatical information. 71.2% of respondents use the Russian monolingual dictionary to clarify definitions, 66.7% to find out the correct spelling, 48.5% to search for grammatical information, and 39% to find examples of use. Given that in this question the respondents could choose several variants as an answer, the results also show that the purpose of using different monolingual dictionaries is often very similar. These results suggest that dictionaries are relatively less used to search for grammatical information. Probably, other more convenient tools and sources like *Grammarly* or *Verbformen* are used for this purpose.

3.3.5 Difficulties during the use of dictionaries and possible causes

In the questionnaire, the respondents had to describe the difficulties they encountered when using a dictionary and had to suggest some possible causes of these difficulties. The most common answer (83.3% of respondents) is that the word or word equivalent searched is not included in the dictionary. Respondents have indicated that often they could not find some specific information (e.g., on domain-specific terms – 61.5%), while the third most frequently mentioned difficulty is that the definition or explanation provided is not clear to the user (38.5%). Possible reasons often include the inability to find the information needed and the lack of experience in searching for such information, as well as the lack of up-to-date information, even in electronic dictionaries. The responses also indicate that some very specific information, such as domain-specific terms, is not included even in modern dictionaries. The same applies to the cases when explanations or equivalents of archaisms are being searched for. Respondents also mention unclear definitions, as well as the fact that dictionaries provide several equivalents in the target language without additional explanation, making it difficult to choose the appropriate translation variant.

Respondents indicate that sometimes dictionaries do not provide explanations of the abbreviations and clarifications used by the compilers. This is the case, for example, of the previously-mentioned resource *Letonika* (https://www.letonika.lv/), which contains several bilingual translation dictionaries with the Latvian language. Another fact deduced from the respondents' answers is important for foreign language teaching. Namely, 5 respondents (representing 6.4% of respondents) mention that they do not know the alphabet of the language of the dictionary, which hinders a quick search in a printed dictionary, which is organized alphabetically. This is a surprising response from translation studies students, but it could be useful for foreign language teachers in shaping the content of their curriculum. This fact illustrates the consequences of a careless attitude toward teaching and learning the alphabet in foreign language courses. Knowledge of the alphabet is a basic condition for

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

the use of printed alphabetical dictionaries, and its acquisition is useful for further studies, not only for a meaningful use of dictionaries and practical translation.

3.3.6 Evaluation of dictionaries and use of other lexicographic materials in translation studies

The following section of the questionnaire surveys how students rate their most commonly used dictionaries, as well as what other reference materials, especially electronic tools, they use in their translation studies. Here are some of the key findings from students in the 2nd-year to 4th-year courses at VUAS FTS that have already mastered the basic skills of dictionary use and gained insight into the basic questions of lexicography theory. Dictionaries that are used by respondents and are available in the foreign language and translation courses were evaluated according to the criteria developed by Patricia Glowania. Glowania emphasizes that, in evaluating dictionaries, attention must be paid to aspects such as the vocabulary included in the dictionary, the indications of the word class, the explanations of the meanings given in the dictionary and their comprehensibility, as well as the microstructure of the dictionary (Glowania 2014, p. 13). It should be noted that this does not take into account specialised dictionaries or dictionaries of domain-specific terms, where the approach of evaluation is different, including some additional aspects. Although the subjective aspect of the respondents' assessment is not excluded (as they are not yet professionals in the field), these results provide some insight into the dictionaries already employed by translation studies students.

When respondents were asked to indicate only three specific dictionaries that they found particularly useful and would recommend to others, different e-dictionaries were among the most mentioned. 52.56% of the respondents named various electronic dictionaries and websites with several dictionaries as the most useful and most frequently used, as is the case of the consolidated website *Tēzaurs* (https://tezaurs.lv/), which comprises 333 different sources. The next most frequently mentioned recommendation (appearing 35 times in the answers of the respondents) is the digital resource *Letonika* (https://www.letonika.lv/), which contains dictionaries of different types and language combinations. In total, 18 respondents characterized as very useful the Latvian National Terminology Portal (*Latvijas Nacionālais terminoloģijas portals*, https://termini.gov.lv/), which includes collections of various terms in different language combinations. The *Cambridge Dictionary* website (https://dictionary.cambridge.org/) and the *Merriam-Webster Dictionary* (https://www.merriam-webster.com/), mentioned 16 and 10 times respectively, are also considered to be useful and valuable. *Duden's* electronic dictionary in German (https://www.duden.de/) is also mentioned by 10 respondents. In addition to these frequently mentioned sources, several other electronic dictionaries are named in the questionnaire less than 10 times: e. g., the dictionaries offered by the Latvian language technology company *Tilde* (6 suggestions in the respondents' answers), which are mostly paid products. The *Digitales Wörterbuch der Deutschen Sprache* (*DWDS*, https://www.dwds.de/) is also recommended 6 times.

Despite the common argument that no one will use printed or paper dictionaries in the 21st century, the respondents indicated the significant use of printed dictionaries from different publishing houses and years of publication. These are bilingual translation dictionaries in all language combinations mentioned in the survey: English-Latvian-English, German-Latvian-German, or Russian-Latvian-Russian, as is the case of the bilingual translation dictionaries published by *Avots*. Some questionnaires indicate both the name of the dictionary and

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

its compiler, for example, *Dzintra Kalniņa's Latvian-English, English-Latvian Dictionary* or *Dictionary of the Russian Language by Ozhegov*. The monolingual German dictionaries of the *Duden* and *Langescheidt* publishing houses are also mentioned, as well as special dictionaries, such as *Svešvārdu vārdnīca* (*The Dictionary of Foreign Words* by J. Baldunčiks, *Jumava* Publishing House). A total of 14 different printed dictionaries are recommended in the questionnaires. It should be noted here that respondents who are intermediate level translation students pay attention to the compiler and publisher of the dictionary, who become the authorities if the dictionary is good and useful for translation studies. This information is important to assess the reliability of the dictionary in the translators' future practice. These examples of dictionaries suggested by students allow concluding that students use both electronic and qualitative printed dictionaries. Undoubtedly, the share of electronic dictionaries is higher, but the number of dictionaries of certain language combinations depends on the working languages of the respective respondents and the frequency of use of these languages within translation studies.

3.3.7 Difficulties in the use of dictionaries: shortcomings and suggestions for their improvement

This section describes the advantages of the dictionaries mentioned in the questionnaires, their specific shortcomings and recommendations for their improvement. The most frequently mentioned shortcoming is that a dictionary does not contain the required information, and does not include a specific word, its translations or explanations. Unfortunately, the purposes of using dictionaries can be very different, so it is difficult to deduce from the respondents' answers exactly what type of information they have been looking for. Perhaps this shortcoming could be overcome by implementing a function in electronic dictionaries that counts, records and collects all words entered in the search box (and searched more frequently), which would allow compilers to quickly supplement the content of dictionaries with words entered as queries by users but not found. The possibility of sending user comments to the dictionary authors could also address this shortcoming. Students of translation studies would certainly be interested in improving and updating the content of dictionaries. Such an initiative is also recommended to students during the VUAS lexicography study course.

Another disadvantage is also the relatively high price of subscribing to an online dictionary, such as *Letonika*, which, according to the survey data, is one of the lexicographic resources most frequently used. Although this resource is often subscribed to by educational institutions, its use after studies should be considered a paid one.

It has also been found that users find the layout of some dictionaries to be confusing and inconvenient. This is indicated regarding both printed and electronic dictionaries (which, of course, have a much more complex structure). The resource *Letonika*, already discussed above, deserves several more critical remarks in the respondents' assessment, as it does not contain explanations of the abbreviations used in the dictionary (which, for example, in the Latvian-German-Latvian dictionary are given in German, as the metalanguage of the dictionary is German). Students do not always know what the abbreviations 'f', 'm' or 'vi' and 'vt' mean, but the explanation is not included in this resource and cannot be found at all. Respondents believe that the explanation of these dictionary metalanguage abbreviations should be given by each word where the abbreviation is used. This would make it much easier to use the dictionary. On the other hand, the dictionary is a tool that requires a min-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

imum of linguistic knowledge, and university students should know what these abbreviations mean. Access to the Cyrillic alphabet would be very useful for the Russian-Latvian and Latvian-Russian dictionaries available in *Letonika*. A transliteration table would certainly be efficient, as it would allow entering the searched item in the search box of the electronic dictionary faster since Russian characters are not always available on computer keyboards. Another shortcoming of the dictionaries available in the resource *Letonika* is that the German-Latvian dictionary searches only for the basic forms of words, but if the noun is written in the plural or the past participle of the verb (in German – *Partizip II*), then no match is obtained. However, in other electronic dictionaries, such as *LEO* (https://dict.leo.org/german-english/), words and their translation can be found in any searchable form. An additional advantage of the *LEO* dictionary over the *Letonika* is the audio playback feature, which provides access to the pronunciation of the word for better learning and use. Unfortunately, the *LEO* dictionary does not include the Latvian language.

The type of printed paper dictionaries is often cited as a shortcoming in the questionnaire, as many printed dictionaries (particularly special dictionaries from fields such as construction, machinery or medicine) are still not available in a digital format. Although respondents admit that printed dictionaries sometimes provide more useful information in terms of contents and the information provided is clear, such dictionaries are heavy and inconvenient in the mobile era. They require spending more time to find the information needed and therefore the translation process is less time-efficient. Respondents prefer using sites that combine several dictionaries (e. g., *Tēzaurs, DWDS*), so that each dictionary does not have to be opened and the search query entered separately. As the time-saving aspect of dictionary use is relevant, dictionaries that provide other useful information, such as grammar overview, noun declension or verb conjugation tables and other reference materials, are preferred because there is no need to search for such information in another resource.

4. Conclusions

This article describes a study that analyses the habits of 78 translation students regarding their use of dictionaries, as it is especially important in the field of translation to be able to use these tools, which have a direct impact on the performance of translators. The results of the study are significant because the assessment of Latvian students' dictionary use habits and skills has not previously been a subject of research. Moreover, the obtained data can subsequently be compared with the experience of other universities, both in Latvia and in other countries. The analysis of translation students' habits helps dictionary compilers to adapt lexicographic publications to the needs of today's generation of users, as well as assess the drawbacks of existing dictionaries.

Although the questionnaire does not include the question suggested by R. R. K. Hartmann (2000, p. 387) about the first dictionary owned by respondents, VUAS TSF students address this question as part of a creative assignment during the course *Terminology and Lexicography*. During the said course, after becoming acquainted with general theoretical aspects, students present an analysis of their first dictionary. According to the author's observations, when working on this assignment students perceive the dictionary completely differently, as they understand the meaning of a dictionary analysis and evaluation through the so-called "I prism". The word *dictionary* is also very important in the 21st century when the fast, but not always reliable, *Google Translate* tool is available. It is worth emphasising once again the pedagogical importance of developing dictionary-using skills among students.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Altogether, the results mirror the findings of other researchers in the field of dictionary use (e. g., Atkins/Varantola 1998; Hamouda 2013; Lew/Galas 2008), confirming the assumption that students do not read dictionary prefaces and users' guides (being therefore unable to properly navigate the dictionary), that they do not take full advantage of the dictionary, that they do not know how to find the required information if several options are provided, etc. Students should be aware that printed dictionaries also need to be consulted occasionally, for example when they cannot find an equivalent in Latvian. This was the case in the practical terminology course with the search for the English word *mural* when the Latvian version was found in the English-Latvian translating dictionary of *Letonika* and also in Jūlijs Anderson's *Mākslas un kultūras vārdnīca* (Dictionary of Art and Culture (Zvaigzne ABC, 2011)).

The results of the research show that VUAS TSF students also need in-depth training on the use of dictionaries, especially considering the enormous changes as a consequence of the technological advances of the 21st century, as well as the rapid development and transformation of various fields, including lexicographic practices. Extensive and in-depth training of potential dictionary users to optimize the use of both printed and electronic dictionaries is needed to be more intensive than before, and the results of the study indicate that there are still many recommendations for improving existing dictionaries with the Latvian language. Similarly, translation courses should evaluate a variety of electronic dictionaries and other electronic resources (Kodura 2016, p. 235). For example, translation courses should also address the evaluation of various electronic resources, such as the online encyclopaedia *Latvijas daba* (https://www.latvijasdaba.lv), environmental dictionary *EnDic2004* (https://mot.kielikone.fi/mot/endic/netmot.exe?UI=ened&height=165), professional translators' forums (https://www.proz.com, https://www.translatorscafe.com, https://translatorforum.de), the origin and author of the resource consulted, reliability, country domain, etc. These will be the topics for further research.

It should be noted that this study is a compilation of data analysis of students from a single university and the results can not be generalised. However, the results obtained in this study provide some insight, especially when compared to similar studies in analogous groups of respondents in Latvia and other countries. The results of this research can supplement the current knowledge about the profile of dictionary users, as well as which skills translation students still need to develop, what shortcomings have been identified in the existing dictionaries with the Latvian language, and which difficulties in usage and possible solutions have been analysed. The study also provides an insight into the current situation to promote the development of lexicographical culture in translation studies.

References

Atkins, B. T. S./Varantola, K. (1998): Monitoring dictionary use. In: Atkins, B. T. S. (ed.): Using dictionaries. Tübingen, pp. 83–122.

Dringó-Horváth, I. (2011): Typen und Untypen elektronischer Wörterbücher. In: Haase, M./Masát, A. (eds.): Jahrbuch der ungarischen Germanistik 2010. Budapest/Bonn, pp. 67–88.

Dringó-Horváth, I. (2012): Lernstrategien im Umgang mit digitalen Wörterbüchern. Themenheft Lernstrategien. Fremdsprache Deutsch 46/2012, pp. 34–40.

Dringó-Horváth, I. (2014): Wörterbuchdidaktik für digitale Wörterbücher. In: Dringó-Horváth, I./Fülöp, J./Hollós, Z./Szatmári, P./Szentpétery-Czeglédy, A./Zakariás, E. (eds.): Das Wort – ein weites Feld: Festschrift für Regina Hessky. Budapest, pp. 218–228.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Glowania, P. (2014): Wortschatzarbeit im DaZ-Unterricht. Wörterbücher im Vergleich. Studienarbeit. w. p.

Hamouda, A. (2013): A study of dictionary use by Saudi EFL students at Qassim University. In Study in English Language Teaching 1 (1), pp. 227–257.

Hartmann, R. R. K. (1987): Four perspectives on dictionary use: a critical review of research methods. In: Cowie, A. P. (ed.): The dictionary and the language learner. Papers from the EURALEX Seminar at the University of Leeds, 1–3 April 1985. (= Lexicographica Series Maior 17). Tübingen, pp. 11–28.

Hartmann, R. R. K. (1999): Case study: the Exeter University Survey of Dictionary Use. In: Hartmann, R. R. K. (ed.): Dictionaries and language learning: recommendations, national reports and thematics reports from The TNT Sub-Project 9: Dictionaries. Berlin.

Hartmann, R. R. K. (2000): European dictionary culture. The Exeter case study of dictionary use among university students, against the wider context of the reports and reccomendations of the thematic network project in the area of languages (1996–1999). In: Heid, U./Evert, S./Lehmann, E./Rohter, Ch. (eds.): Proceedings oft he Ninth EURALEX International Congress, EURALEX 2000, Stuttgart, Germany. Stuttgart, pp. 385–391.

Hessky, R. (2009): Wortschatzarbeit – mit oder ohne Wörterbuch. In: Field-Knapp, I. (ed.): Deutsch als Fremdsprache. Sprachdidaktische Überlegungen zu Wortschatz und Textkompetenz. Budapest, pp. 12–22.

Kodura, M. (2016): Dictionary-using skills of translation students. Społeczeństwo. Edukacja. Język, 4, Panstwowa Wyzsza Szkola Zawodowa w Plocku, pp. 235–242.

Lew, R./Galas, K. (2008): Can dictionary skills be taught? The effectiveness of lexicographic training for primary-school-level Polish learners of English. In: Bernal, E./DeCesaris, J. (eds.): Proceedings of the XIII EURALEX International Congress. Barcelona, pp. 1273–1285.

Lew, R. (2004): Which dictionary for whom? Receptive use of bilingual, monolingual and semi-bilingual dictionaries by Polish learners of English. Motovex.

Nesi, H. (2013): Researching users and use of dictionaries. In: Jackson, H. (ed.): The Bloomsbury companion to lexicography. London, pp. 62–74.

Paradowska, U. (2020): Web-based resources and web searching skills for translators with a specific focus on the Polish-English language pair.
http://www.cttl.org/uploads/5/2/4/3/5243866/cttl_e_2020_6_urszula_paradowska.pdf (last access: 12-03-2022).

Sviķe, S. (in prep.): Vārdnīcu lietošana otrās un trešās svešvalodas apguvē: mācīšanās stratēģija / Using dictionaries in second and third language learning: The learning strategy. In: Laiveniece, D. (ed.): Valodu apguve: problēmas un perspektīva. Liepāja, pp. X–X.

Świątkiewicz-Siklucka, D. (2008): Zur Wörterbuchbenutzung im Fremdsprachenunterricht / beim Fremdsprachenunterricht. In: Rezgui, S./Ingrund, B. (eds.): Wörterbücher und Nachschlagewerke im Sprachunterricht. Babylonia, Nr. 4, pp. 28–32.

Taudic, D./Krause, A. (2017): European Master's in translation competence framework 2017.
https://ec.europa.eu/info/sites/default/files/emt_competence_fwk_2017_en_web.pdf (last access: 06-05-2022).

The standard for the profession of translator (*Tulkotāja profesijas standarts*) (2012): Agreed at the meeting of the Vocational Education and Employment Tripartite Cooperation Subcouncil on February 15, 2012, Minutes No. 2.
https://registri.visc.gov.lv/profizglitiba/dokumenti/standarti/ps0102.pdf (last access: 10-03-2022).

Tono, Y. (2012): Research on dictionary use in the context of foreign language learning: focus on reading comprehension. Tübingen.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

VISC (2020): Valsts Izglītība satura centrs (National Centre for Education Republic of Latvia). Svešvaloda (vācu valoda); Svešvaloda I (vācu valoda): pamatkursu programmas paraugs vispārējai vidējai izglītībai. (Foreign Language (German); Foreign Language I (German): sample programme for general upper-secondary education). I. Baumane, A. Jonasta, K. Kugrēna, I. Rorbaha, Valsts Izglītības satura centrs (VISC). [Rīga]: Valsts Izglītības satura centrs.

Contact information

Silga Sviķe
Ventspils University of Applied Sciences
silga.svike@venta.lv

Acknowledgements

This research has been funded by the Latvian Council of Science, project "Smart complex of information systems of specialized biology lexis for the research and preservation of linguistic diversity", No. lzp-2020/1-0179

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Carole Tiberius/Jelena Kallas/Svetla Koeva/
Margit Langemets/Iztok Kosem

AN INSIGHT INTO LEXICOGRAPHIC PRACTICES IN EUROPE

Results of the extended ELEXIS Survey on User Needs

Abstract The paper presents the results of a survey on lexicographic practices and lexicographers' needs across Europe that was conducted in the context of the Horizon 2020 project European Lexicographic Infrastructure (ELEXIS) among the observer institutions of the project. The survey is a revised and upgraded version of the survey which was originally conducted among ELEXIS lexicographic partner institutions in 2018 (Kallas et al. 2019a). The main goal of this new survey was to complement the data from the ELEXIS lexicographic partner institutions in order to get a more complete picture of lexicographic practices both for born-digital and retro-digitised resources in Europe. The results offer a detailed insight into many aspects of the lexicographic process at European institutions, such as funding, training, staff, lexicographic expertise, software and tools. In addition, the survey reflects on current trends in lexicography and reveals what institutions see as the most important emerging trends that will affect lexicography in the short-term and long-term future. Overall, the results provide valuable input informing the development of tools, resources, guidelines and training materials within ELEXIS.

Keywords E-lexicography; lexicographic practices; lexicographers' needs; survey; ELEXIS

1. Introduction

In each and every European country, elaborate efforts are put into the development of lexicographic resources describing the language(s) of the community. Although confronted with similar problems, cooperation on a larger European scale has long been limited. The result is a rather heterogeneous lexicographic landscape characterised, on the one hand, by stand-alone lexicographic resources, and, on the other hand, by a significant variation in the level of expertise and resources available. Furthermore, as noted by Leroyer and Køhler Simonsen (2020, p. 184) „the digital revolution […] is leading to metamorphoses not only in dictionary making processes and dictionary forms, but also in dictionary use and in the general status of lexicography". The field finds itself in a transitional phase and as yet there is little consensus on the way forward (Rundell 2015, p. 310). Addressing these issues and paving the way for future lexicography, is precisely the goal of the Horizon 2020 ELEXIS[1] project, which is dedicated to creating a sustainable infrastructure for lexicography (Krek et al. 2018, 2019; Pedersen et al. 2018; Woldrich et al. 2020).

To gain more insight into current lexicographic practices, workflows and the specific needs of lexicographers, a number of surveys have been conducted within the project. This paper presents the results of the latest survey which was targeted at the ELEXIS observer institutions. The main goal of this survey was to complement the data from the earlier surveys (Kallas et al. 2019a, b) in order to get a more complete picture of lexicographic practices both for born-digital and retrodigitised resources in different institutions in Europe.

[1] https://elex.is/ (last access: 25-03-2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

After setting the background and introducing the methodology, we will discuss the results specifically focussing on similarities and differences between the answers from the observer institutions and those from the lexicographic partner institutions (Kallas et al. 2019a).

2. Background and methodology

2.1 ELEXIS

The main objective of ELEXIS is to create a sustainable infrastructure for lexicography to 1) enable efficient access to high quality lexicographic data so that it can also be used by other fields including Natural Language Processing (NLP), artificial intelligence (AI) and digital humanities, and 2) bridge the gap between more advanced and less-resourced scholarly communities working on lexicographic resources. To realise these goals, ELEXIS has an inclusive multi-layered organisation that aims at engaging different user groups with various levels of intensity during the project (see Fig. 1). The core of the organisational structure consists of 17 consortium partners. The consortium is composed of content-holding institutions and researchers with complementary backgrounds: lexicography, digital humanities, standardisation, language technology, Semantic Web and AI. Furthermore, the consortium cooperates with existing infrastructures, i. e. CLARIN and DARIAH.

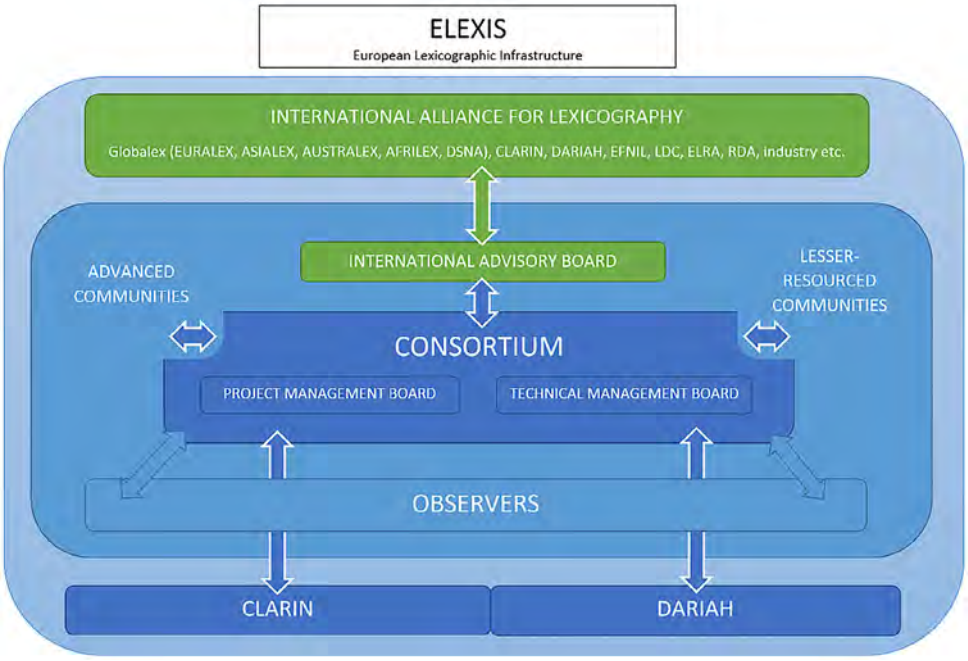


Fig. 1: ELEXIS organisational structure

Another organisational layer is formed by observer institutions that are directly included in outreach and dissemination activities through various channels. The central group of institutions that fall under the observer category are typically, but not exclusively, those producing quality lexicographic data and resources. As of March 2022, ELEXIS has 56 observers.[2]

[2] https://elex.is/observers/ (last access: 25-03-2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Many of the ELEXIS partners and observers already participated and collaborated in the European Network of e-Lexicography (ENeL) COST action[3] (2013–2017), which brought together the lexicographic community in Europe on a larger scale for the first time. In order to learn more about the lexicographic community, a number of surveys have been carried out within COST ENeL, providing valuable information on various aspects of the lexicographic workflow (Tiberius/Krek 2014), the tools that are used (Krek et al. 2014), and on the automation of the lexicographic process (Tiberius et al. 2015).

ELEXIS has used the results from the COST action. However, due to rapid changes in the field, further research and updates were needed. In July 2018, two surveys were launched focussing on lexicographers‘ needs. The first survey was targeted specifically at individual lexicographers, the second survey (which was more comprehensive) focussed on institutions and was sent to the eleven ELEXIS lexicographic partner institutions.[4] In the final months of 2019, a third survey was held among partner and observer institutions in order to gain an insight into their licensing practices (Kosem et al. 2021). In May 2020, an impact survey was launched to assess different aspects of the technical and social infrastructure ELEXIS provides (Wissik et al. 2020).

As the 2018 survey on lexicographers‘ needs for institutions was only sent to the ELEXIS lexicographic partner institutions, it was decided to collect more data by extending this survey to the observer institutions. With this extra data, we get a more complete picture of lexicographic practices across Europe, different tools and methods used by lexicographers, as well as the lexicographic needs that institutions in Europe have now or anticipate to have in the short-term and long-term future.

2.2 Methodology

The survey for observer institutions is a revised and upgraded version of the survey which was originally conducted among the lexicographic partner institutions in 2018. The method chosen for the survey was an online questionnaire. The earlier surveys were conducted in Google Forms as the tool was easy to use and administer. However, as Google Forms does not support nesting of questions, which led to some unexpected results during the analysis, it was decided to switch to a more advanced survey system, i. e. 1ka.[5] Furthermore, we improved the wording of certain questions, which were either unclear or interpreted in different ways by the respondents in the earlier surveys. For instance, the question ‘Do you outsource parts of your lexicographic work to an IT company or language technology company?’ was replaced by ‘Does your institution use services of external providers, such as IT companies, language technology companies, self-employed software developers?’ as the term ‘outsourced’ was not understood in the same way by all respondents.

The survey for observer institutions was the longest survey so far. It contained 121 questions divided into 6 sections: 1) General information, 2) Types of lexicographic resources.

[3] https://www.elexicography.eu/ (last access: 25-03-2022).

[4] The ELEXIS lexicographic partner institutions are the Austrian Academy of Sciences, Institute for Bulgarian Language Prof Lyubomir Andreychin, Society for Danish Language and Literature, Institute of the Estonian Language, Trier University, Trier Center for Digital Humanities, Hungarian Academy of Sciences, Research Institute for Linguistics, K Dictionaries Ltd, Dutch Language Institute, Belgrade Center for Digital Humanities, Jožef Stefan Institute, and the Real Academia Española.

[5] https://www.1ka.si/d/en (last access: 25-03-2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Software and tools supporting the workflow, 3) Publication and access. Crowdsourcing and Gamification, 4) Retrodigitised dictionaries, 5) Data formats. Metadata. Availability, 6) Past and Future. To obtain as much information as possible, the survey included not only "yes/no" questions, and multiple choice questions, but also many open-ended questions. Not all questions were obligatory. The intention was that one survey would be completed per institution and that it would be completed by a representative on behalf of the institution. We cannot, however, exclude that some personal opinions are reflected in some of the answers given.

The survey was opened from 13 July 2020 till 9 November 2021 to allow as many observer institutions as possible to complete it. Towards the end of this period personalised reminders were sent out.

3. Results

In this section, we present an analysis of some of the main results, pointing out si-milarities and differences between the ELEXIS observer and the lexicographic partner institutions (Kallas et al. 2019a). Findings from other surveys are included when relevant. Overall, the response rate was quite high. The survey was completed by 54 observer institutions from 32 countries.

3.1 Respondents' background, institutions and projects

The results show that the representatives of the institutions completing the survey are primarily people working as a corpus linguist, computational lexicographer or computational linguist, having at least 6 years of experience in lexicography (the majority having in between 11–20 years of experience in the field), and holding a PhD mostly in language or linguistics. This is similar to the results from the partner institutions, except that the representatives from the partner institutions have been in the field even longer, the majority having more than 20 years of experience in lexicography.

As can be seen in Table 1, there were slightly more universities than public institutions among the responding observer institutions. This is different from the situation among the lexicographic partner institutions which are mostly public institutions or non-profit organisations. To date there are no private/commercial companies among the observers.

Type of organisation	
Public institution (eg. National Institute, National Centre or Society)	23
University or one of its departments (usually legal person in public law)	26
Private / commercial company	0
Non-profit organisation (NGO)	4
Mixture of public and private (public-private partnership, PPP)	1

Table 1: Types of organisation of the observer institutions

The majority of the observer institutions receive funding for their lexicographic work at the national level. Some are directly funded by the government, others rely on grants from

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

national research agencies.[6] Seven institutions indicated receiving private funding of which three (university and non-profit) rely solely on private funding.

These observations seem to suggest that lexicographic projects in Europe are heavily dependent on national funding by the government. This is in line with earlier results (Kallas et al. 2019a, p. 57) which suggested that lexicographic work in Europe is mainly done in public institutions and non-profit organisations. It is also corresponds with the findings of the European survey on dictionary use and culture (Kosem et al. 2019, p. 96) where it was reported that in the majority of the countries participating in the survey, monolingual dictionaries are published solely or mainly by public institutions funded by the government. This observation is further confirmed by the answers on lexicographic expertise which show that monolingual projects are primarily carried out by public institutions, whereas bilingual and multilingual projects are mentioned more frequently by universities.

Similar to the ELEXIS partner institutions, the majority of the observer institutions employ between 1–10 lexicographers (summed up into full-time employment). Note though that 7 institutions do not employ any lexicographers at all. This is not completely unexpected as lexicographic work per se is not always a core task of the observer institutions.

Most lexicographers at the observer institutions do not work exclusively on lexicographic projects and spend more than 50% of their time on other tasks such as teaching, project management, and public relations. Only four institutions indicated that their lexicographers work exclusively on lexicographic projects. Three of those employ between 11–25 FTE lexicographers. Unlike the partner institutions, only less than half of the observer institutions provide training for their lexicographers. If training is provided, in-house training is most common, followed by specialised workshops and training schools. This is again in line with the findings from the earlier surveys where we observed that specific lexicographic training is often received on the job rather than obtained through formal education programmes. These findings emphasise the importance of the ELEXIS curriculum (Tasovac et al. 2022) and degree programmes such as EMLex (European Master in Lexicography)[7] for the training of young generations of lexicographers.

Most observer institutions do have IT support, although 11 indicated that they have no software developers/IT people working at the institution. All partner institutions indicated having IT support. However, in most cases, the IT people do not work full-time on lexicographic projects. At most observer institutions they spend even less than 10% of their time on lexicographic projects.[8]

About half of the observer institutions indicated that they do use services of external providers, such as IT companies, language technology companies, self-employed software developers (mostly sporadically and some regularly). Four institutions indicated that they do not use external providers at the moment, but that they are planning to do so in the future. The remaining institutions indicated that they do not use such services at all. Development of an online application and user interface is the task which is most commonly outsourced. This is in line with the results from the survey from the partner institutions.

[6] One institution noted that their National Science Foundation will only finance lexicographical projects if dictionary making is included in some linguistic topic, as then it is considered as a science project.

[7] https://www.emlex.phil.fau.eu/ (last access: 25-03-2022).

[8] We do not know this for the lexicographic partner institutions as this question was not included in their version of the survey.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Next are retrodigitisation tasks such as scanning, typing and conversion. Retrodigitisation scores slightly higher in the survey for the observers than in the survey for the partner institutions. Other tasks which are commonly outsourced are the development of a Dictionary Writing System (DWS) or a Corpus Query System (CQS), setting up a database or the creation of a mobile app.

Like the partner institutions, the observer institutions have a 'varied' lexicographic expertise ranging from general and terminological dictionaries to specialised, learner's, historical, and dialect dictionaries (see Fig. 2).

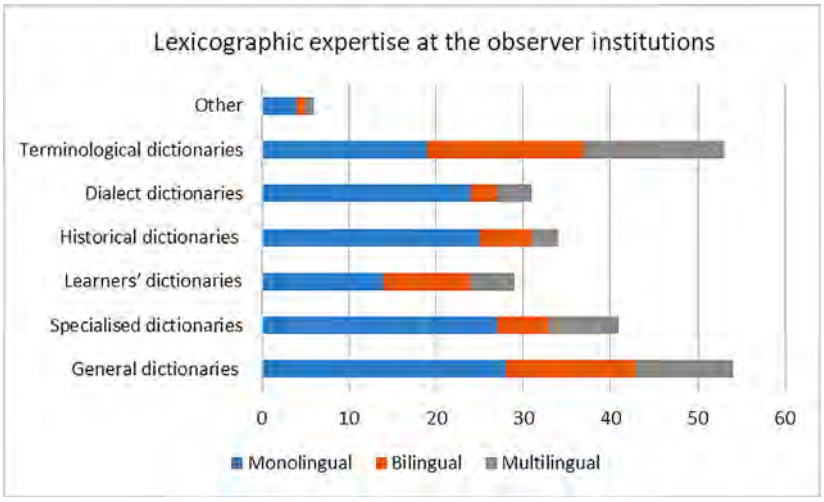


Fig. 2: Lexicographic expertise at the observer institutions

Expertise on terminological dictionaries is, however, more represented among the observer institutions than among the lexicographic partner institutions.

The amount of lexicographic resources per institution also differs. The analysis shows that most observer institutions have between 1 and 5 lexicographic resources whereas 5 out of the 11 partner institutions indicated having between 10–50 lexicographic resources and 2 having even more than 50.

The observer institutions were also asked about ongoing and future lexicographic projects. A total of 131 projects were mentioned, mainly specialised, monolingual and bilingual dictionaries. A fair number of IT-based projects, e. g. on automatic term recognition, were also mentioned. Such projects were not mentioned by the lexicographic partner institutions. As for future projects, the results from the observers seem to suggest that there is a shift from the compilation of general dictionaries towards specialised dictionaries focussing e. g. on neologisms, dialects, etymology, multiword expressions or morphology.

Similar to the resources of the partner institutions, most of the lexicographic data from the observer institutions is published online. The number of resources that are published as scanned or photographed electronic dictionaries is, however, much higher in the survey for the observer institutions compared to the survey of the partner institutions, where there was only one institution selecting this option. The main reason for publishing in print is that it is tradition; the dictionary is part of a larger project and previous volumes have appeared in print. This was also the main reason for publishing in print for the partner institutions. These results are also in line with what was reported by Kosem et al. (2019,

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

pp. 109–111) on the status of lexicography (types of dictionaries being compiled and their format) in the 26 countries involved in their study. Other reasons that were mentioned by the observer institutions for publishing in print, are lack of technical support or software and user demand. It was also pointed out that print dictionaries are still convenient when the intended audience does not have other means of accessing the dictionary, i.e. school children or elderly people. Finally, it was noted that printed dictionaries might still be used in the drafting phase for checking.

3.2 Software and tools

During the last decades, there has been a rapid development of Dictionary Writing Systems and Corpus Query Systems moving towards better interoperability between DWS and CQS and, as a next step, integrating them into one tool. The responses to both surveys show that a large number of different commercial, open-source and in-house tools are used to support lexicographic work in Europe. ELEXIS partners mentioned 11 DWSs and 8 CQSs, observers – 26 DWSs[9] and 31 CQSs. Of the various systems, Sketch Engine[10] is the most mentioned CQS and Lexonomy[11] – the most mentioned DWS.

CQSs are used commonly by observer institutions as well as by partner institutions. Of the 52 observer institutions answering these questions, 36 use a CQS and 16 do not (7 of them feel that they need one urgently). Overall, the institutions are satisfied with the CQS they use. The additional wishes expressed overlap with those mentioned by the lexicographic partner institutions, i.e. advanced corpus creation and annotation tools; better metadata management; additional functionalities (e.g. sense clustering; sense annotation and disambiguation; diachronic analysis; detection of translation equivalents); improved user ergonomy and customisation of the user interface according to user profile, e.g. CQS for learners.

For DWSs, the situation is clearly different from that of the partner institutions. Only 21 observers institutions use a DWS while 31 do not (14 of them feel that they need one urgently). In line with earlier results (Tiberius/Krek 2014; Kallas et al. 2019a), we see that at the observer institutions in-house solutions are still very common too. Most of the observer institutions using a DWS seem more or less satisfied with the system they use although concerns are expressed about the long-term sustainability of the system or about keeping up with technical improvements. Reasons mentioned by the observer institutions for not using a DWS are financial difficulties in purchasing lexicographic software or tools, but also the absence of knowledge and technical skills. This is why open-source tools such as Lexonomy are much welcomed. 22 observer institutions mentioned using Lexonomy, in projects (8), for teaching and training (7), and/or for testing (10).

Features of a DWS that are particularly appreciated by the institutions are the availability of support, customisation options, the possibility to adapt and add functionalities, the ability to work with multiple users and real-time updating of the database. Customisation concerns

[9] In the survey a DWS was defined as "a piece of software for writing and producing a dictionary. It might include an editor, a database, a Web interface and various management tools – for allocating work etc. Specialised dictionary editing software includes customisations of existing/standard (XML) editors". We are uncertain if all the systems mentioned fulfil this definition as some seem to be more targeted at terminology or corpus development.

[10] https://www.sketchengine.eu/ (last access: 25-03-2022).

[11] https://lexonomy.elex.is/ (last access: 25-03-2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

mostly schemas, DTDs and menus, search options, and export options (incl. export for saving and transformation (e. g. XML, CSV, JSON, TEI), for printing (e. g. pdf, Indesign), and for publishing online). Additional wishes include easy installation, support for interlinking lexical entries, providing links to corpus examples and metadata, adding multimedia files, and API access. Also the need for publishing policies and licensing regulations was stressed repeatedly. At the time of writing, some of these wishes have already been implemented in Lexonomy, such as adding multimedia files, API access and interlinking.

The integration of CQS and DWS and advanced semi-automatic dictionary drafting were also explicitly expressed wishes. Similar to the results from the survey of partner institutions, a minority can integrate data from the CQS directly into the DWS that they use and very few have them integrated into one system (4 observer institutions and 2 partner institutions). Integration of DWS and CQS has thus not yet become common practice in modern lexicography, although institutions feel that this would be beneficial, especially for the linking, selection and retrieval of examples and collocations.

Quite a few observer institutions note a lack of information regarding usability and effectiveness of available commercial and open-source CQSs and DWSs and mention the availability of documentation and training materials as preliminary requirements for adopting a particular solution. Training and education are part of the ELEXIS agenda and the ELEXIS curriculum (Tasovac et al. 2022) will provide courses on mastering the ELEXIS tools.

3.3 Retrodigitisation

The survey contained a separate section on retrodigitidation, which received more answers from the observer institutions than from the lexicographic partner institutions (25 out of 54 versus 7 out of 11). The dictionaries of great interest for retrodigitisation are again dialectal, historical and onomastic dictionaries. Focus is also on (multi-volume) dictionaries with common vocabulary, which were published in the second half of the twentieth century.

No clear dependence can be found based on the lexicographic tradition in different countries: for example, there are institutions that do not deal with retrodigitisation in both eastern and western parts of Europe and the same applies to institutions working on retrodigitisation. However, a conclusion can be drawn about the type of institution that predominantly deals with retrodigitisation: thirteen public institutions (52%), eight universities or university departments (32%) and four non-profit organisations (16%) reported retrodigitisation, compared to nine public institutions (36%) and fifteen universities (60%) that did not. This suggests that retrodigitisation is more often practised in specialised lexicographic centres than in universities.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

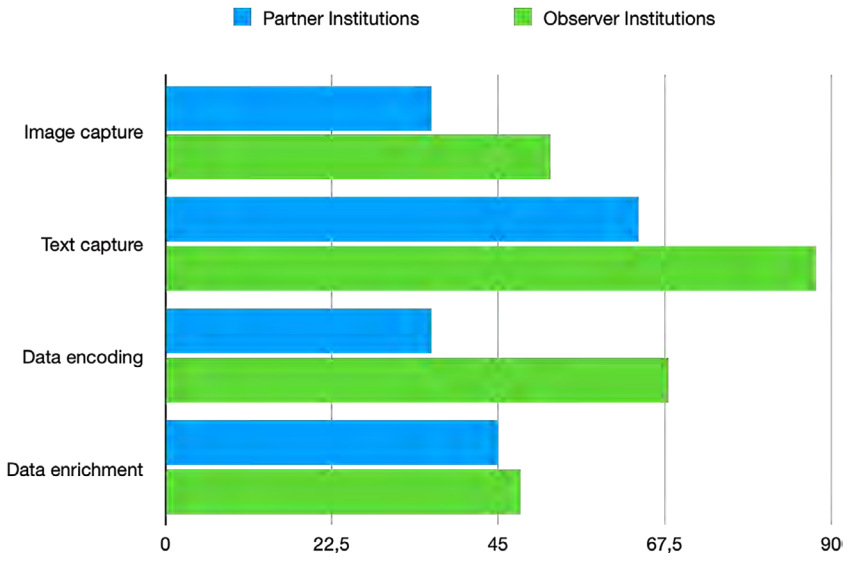


Fig. 3: Phases of retrodigitisation compared in partner and observer institutions

As shown in figure 3, the lexicographers from partner institutions take part mostly in activities such as text capture and data enrichment, while the activities of observer institutions prevail in text capture and data encoding.

The number of institutions (20) that offer access to their retrodigitised resources through an institutional portal or website constitutes 37% of all observer institutions and 80% of those performing retrodigitisation (compared to 45,5% of all partner institutions). Four of the observer institutions (16%) performing retrodigitisation offer access through an API, six (24%) by downloading image files, and 1 (4%) by downloading full text. Among the respondents that took part in this section of the survey fifteen institutions (60%) do not share the full text of their retrodigitised dictionaries with their users, compared with the ten institutions (40%) that reported sharing full text of retrodigitised dictionaries. The reasons for not sharing are copyright restrictions and still ongoing work.

Several options are available with regard to the integration of retrodigitised dictionaries with existing lexicographic resources: eight observer institutions (32%) keep their retrodigitised dictionaries as stand-alone resources, each retrodigitised dictionary has its own website; five observers institutions (20%) have one website or portal which provides access to all retrodigitised dictionaries; three observers institutions (12%) have one website or portal which provides access to all dictionaries (combining retrodigitised and born-digital dictionaries).

In general, interest in retrodigitisation of printed dictionaries is observed in the whole lexicographic community – among the ELEXIS partners and observers. In both surveys similar procedures and software tools were mentioned for the different phases of retrodigitisation (image capture, text capture, data encoding and data enrichment). This is reassuring and suggests that there are already some best practices in place for the retrodigitisation workflow.

4. General observations and wishes for the future

The results from our latest survey show that the lexicographic landscape in Europe is still rather heterogeneous. The observer institutions completing the survey can be divided into three groups, a) under-resourced, b) intermediate, moving towards online, and c) (techno-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

logically) advanced. The ELEXIS infrastructure plays an important role in bridging this gap. For the future, respondents from both the lexicographic partner institutions and the observer institutions would like to see increased interoperability, linking and sharing of resources, more open-source programs and platforms as well as training on how to use them (this is especially important for institutions with limited funding for lexicographic projects, a problem which is mentioned frequently by the respondents), more NLP resources for low-resourced languages and (more) stable and established formats for data encoding in lexicographic projects. Although a shift can be observed from non-structured data to structured data, there are still quite a few institutions (46% of the observer institutions and 36% of the lexicographic partner institutions) using non-structured data format (e. g. in Microsoft Word) for at least some of their projects. This is frequently mentioned as a major hurdle for technological advances. ELEXIS aims to overcome this obstacle with the development of a general open standards based framework for internationally interoperable lexicographic work within OASIS.[12]

The respondents also envisage intensive integration of lexicographic data into the Semantic Web, AI, and NLP applications, as well as aggregating stand-alone lexicographic (and also terminological) resources into dictionary portals. Interoperability and linking are also part of the ELEXIS agenda. One of the main results of ELEXIS is the Dictionary Matrix which is formed of extensive links between key elements found in different types of dictionaries – monolingual, multilingual, modern, historical, etc. – creating a universal lexicographic metastructure spanning across languages and time. The Dictionary Matrix will be available as a public service, and the links between dictionary elements will be shared as Linguistic Linked Open Data (LLOD) enabling other fields to exploit the high-quality semantic data from lexicographic resources. To support linking, editing, enriching and publishing data from various sources, a set of services and tools have been developed within ELEXIS dedicated to the conversion of lexicographic resources to a uniform data format (e. g. Elexifier)[13] as well as to the creation of new resources (e. g. Lexonomy).

Considering the obstacles that were mentioned, one of the biggest concerns seems to be funding. The need for funding is voiced in all the ELEXIS surveys and in all parts of Europe, from Ukraine to Iceland, from Portugal to Sweden, although it seems even more urgent in Eastern Europe where the phrase 'lack of funding' tends to be used, whereas in Western Europe the respondents speak of 'difficulties' obtaining funding. In addition, concerns are expressed about the low status of lexicographic work, which forms a constant worry for many institutions.

In line with the results from the lexicographic partner institutions and the individual lexicographers, some observer institutions expressed their concern about the low quality and reliability of (semi-)automatically built resources while high quality lexicographic data is still kept closed under restrictive licenses (both, public institutions and private publishing houses). Within ELEXIS serious efforts have been made to address licensing issues (Boelhouwer et al. 2020) and a number of flexible and diverse licensing options have been identified to encourage contribution of data (or parts of it) to the Dictionary Matrix. The survey results show that the process of making lexical resources more openly available has already started in the lexicographic community. Most partners and observers make their dictionaries available online for free. However, access to the data for reuse by others is still more

[12] https://www.oasis-open.org/committees/tc_home.php?wg_abbrev=lexidma (last access: 25-03-2022).

[13] https://elexifier.elex.is/ (last access: 25-03-2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

restricted. Only a few institutions indicated that their data is available for free without any restrictions. This suggests that more promotion and raising awareness is needed to open up lexicographic data.

A change in the role of lexicographers, as well as a shift in skills, can also be observed. These days, lexicographers are commonly involved in project management, data management, fundraising, teaching, and public relations. Also, there is a shift in the role of lexicographic institutions, as they become more of a data provider and less of a dictionary publisher. One of the ELEXIS goals is precisely to enable reuse of lexicographic data in other fields.

5. Conclusion

The survey of observers has provided further insights into existing practices and needs of lexicographers around Europe. It successfully complements the other surveys conducted in the ELEXIS project and in the ENeL COST action, giving a more detailed overview of the current situation in lexicography, emphasising the need for common standards, open-source tools and comprehensive training materials. It also shows the main wishes, needs and concerns of lexicographic institutions.

The results from all surveys have already provided valuable input for various tasks within the ELEXIS project, and will continue to inspire future developments within the infrastructure. On the basis of the combined results, a lexicographic practice map of Europe can also be devised, which is something we would like to explore in future research.

References

Boelhouwer, B./Kosem, I./Nimb, S./Jakubíček, M./Tiberius, C./Krek, S./Rosenmeier, M. (2020): ELEXIS deliverable 6.2 Recommendations on legal and IPR issues for lexicography. https://elex.is/wp-content/uploads/2020/02/ELEXIS_D6_2_Reccommendations_on_Legal_and_IPR_Issues_for_Lexicography.pdf (last access: 25-03-2022).

Kallas, J./Koeva, S./Kosem, I./Langemets, M./Tiberius, C. (2019a): ELEXIS deliverable 1.1 lexicographic practices in Europe: a survey of user needs. https://elex.is/wp-content/uploads/2020/06/Revised-ELEXIS_D1.1_Lexicographic_Practices_in_Europe_A_Survey_of_User_Needs.pdf (last access: 25-03-2022).

Kallas, J./ Koeva, S./Langemets, M./Tiberius, C./Kosem, I. (2019b): Lexicographic practices in Europe: results of the ELEXIS Survey on User Needs. https://doi.org/10.5281/zenodo.3726841 (last access: 25-03-2022).

Kosem, I./Lew, R./Müller-Spitzer, C./Ribeiro Silveira, M./Wolfer, S. (2019): The image of the monolingual dictionary across Europe. Results of the European survey of dictionary use and culture. In: International Journal of Lexicography 32 (1), pp. 92–114. https://doi.org/10.1093/ijl/ecy022 (last access: 25-03-2022).

Kosem, I./Nimb, S./Tiberius, C./Boelhouwer, B./Krek, S. (2021): License to use: ELEXIS survey on licensing lexicographic data and software. In: Gavriilidou, Z./Mitits, L./Kiosses, S. (eds.): Proceedings of XIX EURALEX Congress: Lexicography for Inclusion, Vol. II, Democritus University of Thrace. https://euralex2020.gr/wp-content/uploads/2021/09/Pages-from-EURALEX2021_ProceedingsBook-Vol2-p705-712.pdf (last access: 25-03-2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Krek, S./Abel, A./Tiberius, C. (2014): Dictionary writing systems & corpus query systems. Survey-COST ENeL-WG3 meeting, February 2015, Vienna. https://www.elexicography.eu/wp-content/uploads/2015/04/LexicographicalWorkflow_DeliverableWG3BolzanoMeeting2014.pdf (last access: 25-03-2022).

Krek, S./McCrae J./Kosem, I./Wissik, T./Tiberius, C./Navigli, R./Pedersen, B. S. (2018): European Lexicographic Infrastructure (ELEXIS). In: Čibej, J./Gorjanc, V./Kosem, I./Krek, S. (eds.): Proceedings of the XVIII EURALEX International Congress on Lexicography in Global Contexts (EURALEX 2018), Ljubljana, Slovenia, 17–21 July 2018, pp. 881–892. http://doi.org/10.5281/zenodo.2599902 (last access: 25-03-2022).

Krek, S./Declerck, T./McCrae, J. P./Wissik, T. (2019): Towards a globallLexicographic infrastructure. In: Adda, G./Choukri, K./ Kasinskaite-Buddeberg, I./Mariani, J./Mazo, H./Sakriani, S. (eds.): Collection of research papers of the 1st International Conference on Language Technologies for All, Paris, 4–6 December 2019, pp. 120–122. http://doi.org/10.5281/zenodo.3607274 (last access: 25-03-2022).

Leroyer, P./Køhler Simonsen, H. (2020): Reconceptualizing lexicography: the broad understanding. In: Gavriilidou, Z/Mitsiaki, M/Fliatouras, A. (eds.): Proceedings of XIX EURALEX Congress: Lexicography for Inclusion, Vol. I, Democritus University of Thrace, pp. 183–192. https://euralex2020.gr/wp-content/uploads/2020/11/EURALEX2020_ProceedingsBook-p183-192.pdf (last access: 25-03-2022).

Pedersen, B. S./McCrae, J./Tiberius, C./Krek, S. (2018): ELEXIS – a European infrastructure fostering cooperation and information exchange among lexicographical research communities. In: Bond, F./Kuribayashi, T./Fellbaum, C./Vossen, P. (eds.): Proceedings of the 9th Global WordNet Conference (GWC 2018), Global Wordnet Association, Singapore, pp. 339–344. http://doi.org/10.5281/zenodo.2599954 (last access: 25-03-2022).

Rundell, M. (2015): From print to digital: implications for dictionary policy and lexicographic conventions. In: Lexikos 25 (1), pp. 301–322. https://www.ajol.info/index.php/lex/article/view/126250 (last access: 25-03-2022).

Tasovac, T./Tiberius, C./Bamberg, C./Bellandi, A./Burch, T./Costa, R./Ďurčo, M./Frontini, F./Hennemann,J./Heylen, K./Jakubíček, M./Khan, F./Klee, A./Kosem, I./Kovář, V./Matuška, O./McCrae, J./Monachini, M./Mörth, K./Munda, T./Quochi, V./Repar, A./Roche, C./Salgado, A./Sievers, H./Váradi, T./Weyand, S./Woldrich, A./Zhanial, S. (2022): D5.3 Overview of online tutorials and instruction manuals. https://elex.is/wp-content/uploads/ELEXIS_D5_3_Overview-of-Online-Tutorials-and-Instruction-Manuals.pdf (last access: 25-03-2022).

Tiberius, C./Krek, S. (2014): Workflow of corpus-based lexicography. Deliverable COST-ENeL-WG3 meeting, July 2014, Bolzano/Bozen. https://www.elexicography.eu/wp-content/uploads/2015/04/LexicographicalWorkflow_DeliverableWG3BolzanoMeeting2014.pdf (last access: 25-03-2022).

Tiberius, C./Heylen, K./Krek, S. (2015): Automatic knowledge acquisition for lexicography. Survey – COST-ENeL-WG3 meeting, August 2015, Herstmonceux. http://www.elexicography.eu/wp-content/uploads/2015/10/ENeL_WG3_Survey-AKA4Lexicography-TiberiusHeylenKrek.pptx (last access: 25-03-2022).

Wissik, T./ Woldrich, A./Kosem, I./Goli, T./Matuška, O. (2020): ELEXIS deliverable D7.7 ELEXIS impact survey. https://elex.is/wp-content/uploads/2020/05/D7.7_ELEXIS_Impact_Survey_final.pdf (last access: 25-03-2022).

Woldrich, A./Goli, T./Kosem, I./Matuška, O./Wissik, T. (2020): ELEXIS: Technical and social infrastructure for lexicography. In: K Lexical News. https://lexicala.com/review/2020/elexis/ (last access: 24-06-2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Contact information

Carole Tiberius
Instituut voor de Nederlandse Taal
carole.tiberius@ivdnt.org

Jelena Kallas
Institute of the Estonian Language
Jelena.kallas@eki.ee

Svetla Koeva
Institute for Bulgarian Language
svetla@dcl.bas.bg

Margit Langements
Institute of the Estonian Language
margit.langemets@eki.ee

Iztok Kosem
Jožef Stefan Institute
iztok.kosem@ijs.si

Acknowledgements

The research received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No 731015.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Agnes Wigestrand Hoftun

CONSULTATION BEHAVIOR IN L1 ERROR CORRECTION

An exploratory study on the use of online resources in the Norwegian context

Abstract This think-aloud study charts the use of online resources by five final-year MA students in a Nordic languages and literacy program based on the analysis of screen and audio recordings of an error-correction task. The article briefly presents some linguistic features of Norwegian Nynorsk that are not common in the context of other European languages, that is, norm optionality with regards to inflection and spelling. While performing the task, the participants were allowed to use all digital aids. This article examines their resource consultation behavior, and it makes use of Laporte/Gilquin's (2018) annotation protocol. The following research questions are posed: What online resources are used by the students? What characterizes the use? Are online resources helpful? This study provides new insights into an as yet little explored topic within the Norwegian context. The findings demonstrate that the participants relied heavily on the official monolingual dictionary Nynorskordboka. Indeed, the dictionary was helpful in the vast majority of the searches, either resulting in error improvement or the validation of a word; that is, many of the searches considered correct words. The findings suggest severe norm insecurity and emphasize the need to improve norm knowledge and metalinguistic knowledge as prerequisites for better utilization of aids. It is also suggested to include necessary information on norm optionality and other commonly queried issues in the dictionary architecture.

Keywords Consultation behavior; L1 error correction; dictionary use; online resources; Norwegian Nynorsk

1. Introduction and background

Despite the growing body of international research on the use and efficiency of online resources, including dictionaries, little is known about this topic in the context of Norway. To the best of the author's knowledge, this article is the first to report on consultation behavior in Norwegian as a first language (L1).

The Norwegian multinorm language situation stands out and affects the development, architecture, and use of language aids. There are two official Norwegian written languages, Bokmål (majority standard) and Nynorsk (minority standard), that have equal official status, and reading and writing in both languages is a part of the Norwegian subject curriculum in schooling (Udir 2013). There is no official pronunciation standard, and dialects are used in all oral contexts (Helset 2021).

Probably the most striking characteristic of Norwegian is that compared to other commonly known European languages, both Bokmål and Nynorsk include significant optionality in terms of spelling and inflection (regarding both single words and whole categories); see section 3. Although the two written standards are mutually intelligible and largely overlap in terms of lexis and syntax (see, for example, Faarlund 2003 and Helset 2021), they differ significantly in terms of morphology and orthography.

Educators are particularly exposed to both norm plurality and optionality in their work, for example, when teaching and evaluating students' writing. Based on the curriculum, teach-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

ers of Norwegian are expected to master both written languages equally well (e. g., University of Stavanger, no date). However, recent research on student teachers' norm competence has indicated they do not have such mastery in Nynorsk (Russdal-Hamre 2020). Using appropriate aids might help in bridging this gap. An online survey has revealed that 82% of the secondary school teachers who participated in the survey use digital dictionaries when correcting students' texts (TNS Gallup/The Norwegian Language Council 2014; Hovdenak/Ims 2016). The twin dictionaries Bokmålsordboka and Nynorskordboka (The Norwegian Language Council/University of Bergen, no date) are the most popular digital dictionaries for Norwegian (Hovdenak/Ims 2016; for a state-of-the art article on monolingual lexicography, see Grønvik et al. 2019). Yet, empirical research on the use of online resources and dictionaries in the context of the Norwegian language is scarce. This article aims to help fill this knowledge gap by focusing on the minority standard Nynorsk and providing insights on users' consultation behavior. The study design does not limit access to online resources and can thus provide insights regarding users' preferences for using other aids.

2. Aim

The aim of this article is to explore the use of online resources and their effect on L1 error correction in Norwegian Nynorsk, that is, to chart what a sample of students who are expected to master both Bokmål and Nynorsk equally well actually do when correcting a student text and what effect their consultations have on their corrections. Following are the research questions.

2.1 What online resources are used by the students?

Based on the survey mentioned above (TNS Gallup/The Norwegian Language Council 2014), it is expected that the participants are familiar with and use the official online dictionary Nynorskordboka. It is also not uncommon to rely on the built-in spell and grammar checker, which is still not well developed for less widely used languages such as Norwegian; for example, it is not able to track or automatically check for norm consistency.

2.2 What characterizes the use?

In accordance with Gilquin/Laporte (2021), individual variation is expected regarding how many searches the participants carry out. Quick and one-tool searches are expected to dominate. Following the user records and dictionary statistics (Rauset 2019; Jansson 2007), it is expected that a considerable number of the searches will concern common and frequent words.

2.3 Are online resources helpful?

In line with the findings of Gilquin/Laporte (2021), Müller-Spitzer et al. (2018), Wolfer et al. (2018), and Wolfer et al. (2016), it is expected that consulting resources will lead to successful error correction in most but not all cases. As shown in Wolfer et al. (2016) and Jansson (2007), being able to detect an error might be a challenge in the first place.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

3. A brief overview of a selection of linguistic features in Norwegian Nynorsk and their presentation in Nynorskordboka[1]

This section very briefly illustrates some examples of norm optionality in terms of inflection and spelling, but it is by no means exhaustive (see, for example, Faarlund 2003; Almenningen/Søyland 2012 and Helset 2021). Screenshots showing relevant sections from the most popular Nynorsk dictionary, Nynorskordboka, are also provided. Consistency of choice is required within a text (see The Norwegian Language Council, no date-a, for more on norm consistency).

3.1 Inflection

Infinitive

The official Nynorsk norm (Almenningen/Søyland 2012) allows no less than three systems: 1) all infinitives ending with -a; 2) all infinitives ending with -e; and 3) a system called *kløyvd infinitiv* (= divided infinitive) where some infinitives must end with -a and the remaining ones end with -e.

The screenshot below shows parts of the dictionary article about the verb to be, *vera*. The inflection pattern opens when the hyperlinked verb label is clicked.

Oppslagsord Ordbokartikkel

vera
vere

IV **vere** verb (norrønt *vera* (eldre *vesa*), *er* (eldre *es*), *verit*)

Bøying i samsvar med gjeldande rettskriving:

vere	Infinitiv	Presens	Preteritum	Presens perfektum	Imperativ
v. kløyvd inf.	å vera å vere	er	var	har vore	ver

Screenshot 1: *Vera* = be (oppslagsord = head word, ordbokartikkel = dictionary article)

Both infinitive suffixes are shown in the head word list and in the inflection table. Clicking on *kløyvd infinitiv* reveals an explanation of the third system, but the hyperlink might be difficult to notice when all the text is blue. Beyond this, no more information on the optionality is provided here. The head word section within the dictionary article contains only the form ending with -e, which is also the type of infinitive used in the example section.

Inflection patterns

Most Norwegian nouns are of one gender only, resulting in one set of suffixes. This also applies to verbs, most of which have one set of tense suffixes and the optional infinitive

[1] A new version of the dictionary has recently been launched. This study was conducted using the previous version of the user interface that was launched in 2009 and is still available.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

suffixes mentioned above. However, several commonly used nouns and verbs have more than one inflection paradigm. The noun *tekst* (= text) can be either masculine or feminine.

Oppslagsord Ordbokartikkel

tekst — **I tekst** m1, f1 (norrønt *textr* m eller *texti* m, frå latin 'vev')

Bøying i samsvar med gjeldande rettskriving:

tekst	Eintal		Fleirtal	
	Ubunden form	Bunden form	Ubunden form	Bunden form
m.	ein tekst	teksten	tekstar	tekstane
f.	ei tekst	teksta	tekster	tekstene

Screenshot 2: *Tekst* = text. Norwegian nouns are inflected in number and definiteness

The verb *bruka* (= use) has no less than three inflection patterns.

Oppslagsord Ordbokartikkel

bruka, bruke — **bruke** v1, v2, v3 (lågtysk *bruken*)

Bøying i samsvar med gjeldande rettskriving:

bruke	Infinitiv	Presens	Preteritum	Presens perfektum	Imperativ
v.	å bruka å bruke	brukar	bruka	har bruka	bruk
v.	å bruka å bruke	bruker	brukte	har brukt	bruk
v.	å bruka å bruke	brukar	brukte	har brukt	bruk

Screenshot 3: *Bruka* = use

The last pattern is a mix of the previous two, taking the present tense form from the first pattern and the past and present perfect forms from the middle one.

Homographs might have different inflection patterns. For example, the verb *føla* can either mean to foal or to feel, each having its own set of suffixes, as indicated by the different codes in the dictionary (respectively v1 and v2).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Oppslagsord	Ordbokartikkel
føla føle	I **føle** v1 få føl; fole, fylje (I) *merra har føla*
føla føle	II **føle** v2 (frå lågtysk)
	1 granske ved å ta på; kjenne (II.4)

Bøying i samsvar med gjeldande rettskriving:

føle	Infinitiv	Presens	Preteritum	Presens perfektum	Imperativ
v1	å føla å føle	følar	føla	har føla	føl

Bøying i samsvar med gjeldande rettskriving:

føle	Infinitiv	Presens	Preteritum	Presens perfektum	Imperativ
v2	å føla å føle	føler	følte	har følt	føl

Screenshot 4: *Føla* = foal/feel

3.2 Spelling

The spelling optionality covers alternative vowels (*lykke*/*lukke* = happiness), consonants (*dobbelmoral*/*dobbeltmoral* = double standard of morality), similar words (*bilde*/*bilete* = picture), and even completely different words/equivalents (*følelse*/*kjensle* = feeling).

Oppslagsord	Ordbokartikkel
lykke lukke	**lykke** f2; el. I **lukke** f2 (norrønt *lukka* og *lykka*, lågtysk (*ge*)*lucke* 'lagnad, lykke')

Screenshot 5: *Lykke*/*lukke* = happiness

The order of appearance of optional forms in the dictionary article is fixed no matter which one is used in the query. No systematic study has been conducted, but it appears that minor spelling differences between optional forms result in one dictionary article. Equivalents with a larger "spelling distance" are presented in separate articles. The one below is not even hyperlinked, although there is a dictionary article for *kjensle*.

Oppslagsord	Ordbokartikkel
følelse	**følelse** m1 kjensle (1-2)

Screenshot 6: *Følelse* = feeling

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

A word that deserves special attention is *ønske* (= wish), as it has a tremendous number of optional forms. The same number of optional forms is not allowed when it is a noun compared to when it is a verb.

Oppslagsord	Ordbokartikkel
ynske ønske	**ynske** n1; el. **ønske** n1 (norrønt *ósk* f, formene med *-n-* kjem av innverknad frå lågtysk *wunsch*) **1** vilje, hug eller lyst til å få eller vinne fram til noko; lengsel etter å oppnå noko *no får du kome fram med ynska dine / ho bar fram ynsket sitt / gå med på alle ynske og krav* • inderleg von eller lengsel (om at det eller det skal skje) *hans høgaste ynske var å kome seg ein tur utanlands* • oppfordring, krav (I) *eit sterkt uttrykt ynske om å få fortgang i saka* **2** venleg, kjærleg tanke, von om noko (for eit anna menneske) *ho hadde følgt dei med ynske og omsut / han kjem med gode ynske til brudgom og brur*
ynskja ynskje ynska ynske ønska ønske ønskja ønskje	II **ynske** v2; el. **ynskje** verb; el. II **ønske** v2; el. **ønskje** verb (norrønt *óskja, ýskja*; formene med *-n-* kjem av innverknad frå lågtysk *wunschen*) **1** ha ynske (I,1) om; kjenne trong, lyst til; gjerne vilje; lengte etter *eg skulle ynskje du slutta med*

Screenshot 7: *Ønske* as the noun wish, as indicated by the code n1, and as the verb wish, indicated by the v2/verb codes

4. Related work

4.1 International studies

Wolfer et al.'s study (2016) on the effectiveness of lexicographic resources obtained data from 78 L1 German students. Their findings suggest that being able to spot a language problem is crucial, and although access to relevant resources enables the highest number of successful revisions, it does not guarantee success.

A multi-method observational study on error correction and the use of online resources (reported on in Müller-Spitzer et al. 2018 and Wolfer et al. 2018) combined task results, screen recordings, and thinking aloud from 43 L2 learners of German. They were to correct 18 unrelated German sentences, each containing one unmarked error, using any aids they

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

wanted. The study showed that L2 learners resort to lexicographic resources to a great extent. The authors identified some important success factors, such as good metalinguistic knowledge and determination to complete the task. The present study and this project have many common features. However, the participants in the present study were working with a coherent piece of writing in order to lay the groundwork for possible norm consistency checks; cf. section 3.

Laporte/Guilquin (2021) studied the use of online resources by 84 L2 learners of English and the effect of those resources in a free composition task. The authors employed screen recordings to study the learners' consultation behavior. The screen recordings were annotated using ELAN software (Wittenburg et al. 2006) and an annotation protocol developed by the researchers themselves (Laporte/Gilquin 2018). They identified a total of 1,543 searches, most of them quick and utilizing one resource. The number of searches per participant varied greatly, from none to 49. The study showed that resource consultation had a positive impact in the majority of cases. The annotation protocol developed by the authors (2018) has been of great use for the present study. Some minor adjustments were necessary to capture the peculiarities of the Norwegian language context, as exemplified in section 3.

4.2 Norwegian studies

Karlsen/Rødningen's (2008) survey of 111 upper-secondary teachers of Norwegian revealed that 86% provide their students with instruction in using aids. However, it is not being given in a systematic way. The aids are mainly used to obtain formal information, such as about spelling, inflection, and optional forms. An important finding is that Nynorsk aids are the most used and that Nynorsk seems to facilitate and prompt instruction in aid use.

Nygaard/Fjeld (2008) examined a sample of unsuccessful searches based on search logs from Bokmålsordboka, the dictionary of the majority written language. Their study shows that misspelling is the main reason when the dictionary does not return any results.

User records and dictionary statistics reveal that users search for commonplace and frequently used words. This indicates that spelling and inflection in a production context are in question (Rauset 2019), which is not typical for L1 dictionaries (Svensén 2009). Rauset (2019) also shows that users open inflection patterns in Nynorskordboka more than twice as often as in Bokmålsordboka and that searches with the twin dictionaries side by side are most common.

The one study on Nynorsk norm competency and speller use (Jansson 2007) showed that secondary school students made fewer spelling and inflection errors when using a speller compared to when writing with no aids, and seven students out of 19 reduced the number of norm deviations by 50% or more. However, the writing of five students did not improve much. Jansson also charted the words that they would have looked up by asking the students to draw a circle around those words while writing without aids. In addition to words that were erroneous, there were many examples of correct commonplace words that the students would have looked up, which is also in line with what the user records for Bokmålsordboka and Nynorskordboka show (Rauset 2019). On the whole, many of the norm deviations were overlooked. Jansson points out that such behavior reflects norm unsteadiness (2007, p. 37).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

5. The present study

The present study is process-oriented; that is, it is mostly concerned with resource consultation and its effect on the language issues rather than the corrected text itself. By combining screen recording and thinking aloud, this study aims to help make the Norwegian user lexicography less of a terra incognita.

5.1 Participants

Five final-year MA students in Nordic languages and literacy studies at the University of Stavanger in Norway volunteered to participate in the study.[2] There are two types of students enrolled and taking the same courses covering the subject content knowledge – future secondary education teachers of Norwegian (a five-year teacher program) and students who have completed a bachelor's degree and are now taking a two-year master's degree. Both groups of students are expected to master both written languages equally well (University of Stavanger, no date).

5.2 The experiment

The participants were asked to verbalize their thoughts and justify their choices during the experiment (Ericsson/Simon 1993); the data from the error-correction task are reported on in this article. The participants were instructed to 1) mark all language errors and 2) propose correct forms using the comment function. They were allowed to use any digital aids they wished.

An authentic text written by a final-year secondary school student was chosen for the error-correction task. It was approximately two and a half pages long (about 1,000 words). A variety of norm deviations were present, (most of) which were elementary and could easily be corrected with the help of an online resource. A coherent piece of writing was given to render possible norm consistency checks. The student text was available in Microsoft Word with the built-in spell and grammar checker turned off.

The author aimed for as natural and relaxed experiment settings as possible. The goal was to elicit the best Nynorsk competence with multiple occasions to reveal what aids the participants normally use and how they use them. The fact that they had unlimited access to aids and no time constraints as such gives reason to assume the participants were able to do their best in regard to norm competence and aids.

5.3 Screen and audio recording

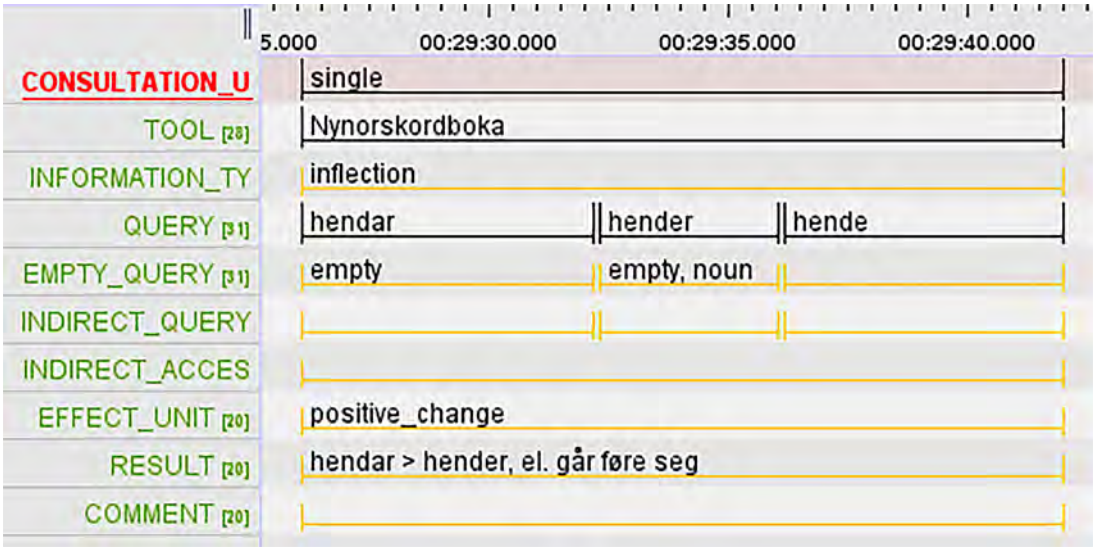
All sessions were recorded using Active Presenter 8. This software allows for the unobtrusive registration of all on-screen actions and audio, which strengthens the ecological validity of this study.

[2] Participants received a 500NOK gift card as compensation.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

5.4 Annotation

The screen and audio recordings were annotated using ELAN software (Wittenburg et al. 2006). All recordings were annotated by the author using Laporte/Gilquin's (2018) annotation protocol developed for research on the use of online resources with minor adjustments. The analysis in this study is based on the tiers (annotation layers) shown on the left-hand side in the screenshot below.



Screenshot 8: Example of coded data on a participant's search (Ask),[3] prompted by *hendar* (= happen, TF,[4] erroneous present tense suffix)

This search (CONSULTATION_UNIT) is annotated as single; that is, it describes continuous use of one tool only. The participant looked up the word *hendar* (= happen, present tense),[5] as annotated in QUERY. The dictionary, Nynorskordboka (annotated in TOOL), did not return any results; cf. EMPTY_QUERY. The participant tried to search for the verb using another present tense suffix. The result of this look-up was also marked as empty, as the dictionary returned the homograph noun hands. On the final try, the participant used one of the infinitive suffixes instead (*hende*), and the dictionary returned the desired verb and its inflection pattern. The participant corrected the erroneous form in the student text (*hendar*) to the form found in the dictionary (*hender*), as annotated in RESULT. Additionally, they proposed a synonym "går føre seg". The kind of information sought or the query intent in this case was inflection, as annotated in INFORMATION_TYPE. The effect of the whole search was positive, and the participant made a change in the text (positive_change on EFFECT_UNIT, as opposed to positive_confirmation when a search validates a form used in the student text). The total resource consultation time for this particular language problem was 16 seconds. See Laporte/Gilquin (2018) for an in-depth description of the annotation protocol. The data were then exported and analyzed in Microsoft Excel.

[3] The participants' names were replaced with other names.

[4] TF stands for text form.

[5] In Nynorsk, there are three (-er/-r, -ar, -Ø) present tense suffixes.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

6. Findings and discussion

A total of 148 consultation units (i. e., searches) were identified. A search may consist of one or several look-ups (and therefore also several queries) addressing one specific language problem.

6.1 What online resources are used by the students?

The participants almost exclusively resorted to the official Nynorsk dictionary, Nynorskordboka, and they expressed, while thinking aloud, that they consider the dictionary a reliable aid that they use often. When looking up a verb with three inflection patterns, one of the participants, Due, stated: "This is why we have the dictionary. It is completely unhuman to have all of it in your head."

Other resources were accessed only three times in total by three participants. Two of these searches can be classified as using other resources, and the latter ended up in Nynorskordboka via a roundabout manner through the Norwegian Language Council's website. Not surprisingly, all three searches concerned two language problems for which the dictionary is not the most suitable resource, passive voice and the use of the determinatives *nokon* and *nokre* (often compared with some and any in English).

Regarding norm optionality and consistency of choice, the find function is especially useful for a time-saving check for norm inconsistency. Only Due and Eir used or expressed that they normally would have used this function, still displaying limited awareness of consistency requirements. None of the participants turned on the built-in spell and grammar checker.

6.2 What characterizes the use?

Single-tool, continuous searches dominated overwhelmingly, as shown in Figure 1. In 125 out of 148 searches, the participants consulted Nynorskordboka and returned to the text to make corrections or moved on. In 115 searches, only one query was carried out. The average time registered on a continuous single-tool search was 22 seconds.[6] As hypothesized, the number of searches made by each participant varied. Due carried out more than double as many searches as Ask, as shown in Figure 1.

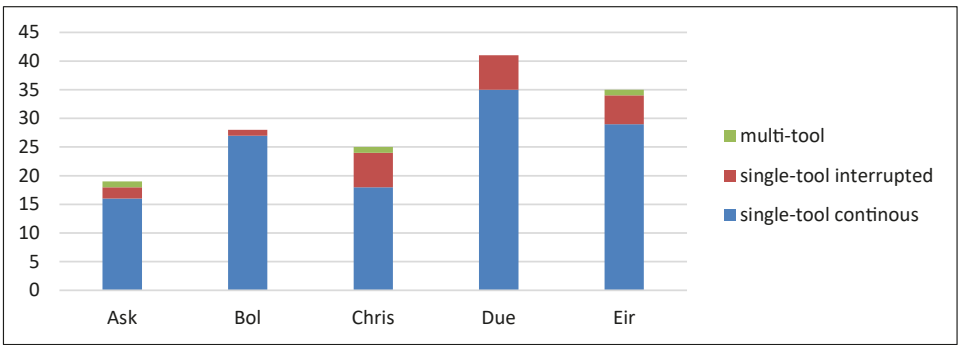


Fig. 1: Number and type of searches per participant

[6] Thinking aloud while performing the task makes the searches somewhat longer.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

The screen recordings also revealed that the participants demonstrated different levels of dictionary skills, for example if and how fast they could locate and extract the information sought. Moreover, the participants showed significant differences in basic information and communication technology (ICT) skills that affected their workflow and therefore also the amount of time they spent on a search. Ask and Due were the most efficient in their consultations, respectively spending 9 and 12 seconds on average for single-tool searches. This seems to cohere with good ICT skills, that is, the fact that they both use keyboard shortcuts to switch between windows and that they are very familiar with the dictionary architecture. Unskillful querying is time-consuming; Eir tended to use word forms other than the base in queries more often than the other participants, which prolonged the searches.

Bol stands out, as this participant spent a longer time on their dictionary consultation than other participants (43 seconds on average for single-tool searches; this number includes the time it took to access the dictionary after a search in Google for each search). This seems to be the result of insufficient dictionary skills, as the participant never opened the needed inflection patterns. Instead, they searched through the example section. The problem with this approach is that the example section in dictionary articles does not present the full range of norm optionality. Following only those word-form choices results in a much narrower norm, that is a norm with fewer optional forms.

The participants seemed to be unaware of other resources that could have helped them resolve language problems other than those concerning spelling and inflection. Ask and Chris were looking for assistance with passive voice (prompted by *giftes*, passive voice of gifte, = marry, erroneous in Nynorsk) in Nynorskordboka. Ask mentioned that googling is useful to see whether others use a given word or expression. Ask skimmed through the snippets in Google and found a website (*Hardanger historielag*) that used "giftast bort" and then returned to the dictionary and got frustrated that it did not help in this case, only showing "a forbidden sign" next to the sought word form.

Screenshot 9: *Giftast* (passive voice of *gifte* = marry and the forbidden sign. When hovering the cursor over it, a text box with this message appears: *Tilslagsord unormert* = word form not standardized. The sign is only present next to the adjective article and not the verb article.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Chris did not get the forbidden sign, as this participant queried the base form (*gifte*). However, Chris was also looking in vain for passive voice in the inflection pattern. Eir did not look for help with *nokon* and *nokre* in Nynorskordboka but instead queried "nokon eller nokre" in Google. Although this participant found The Norwegian Language Council's mini-grammar websites (The Norwegian Language Council, no date-b) that include a section on this topic, they chose to rely on a random Google snippet and based on that corrected *nokon* to the optional, in this case *nokre*, which was an unnecessary correction in the student text. Due also did this, although this participant relied on Nynorskordboka, which only shows both determiners as optional in plural without providing any information on this rather peculiar issue.

Information type

A total of 199 language-related queries were carried out by the participants within the 148 searches. Inflection and spelling constitute the types of information most often sought. Queries regarding grammar and meaning constitute a negligible number,[7] as shown in Figure 2.

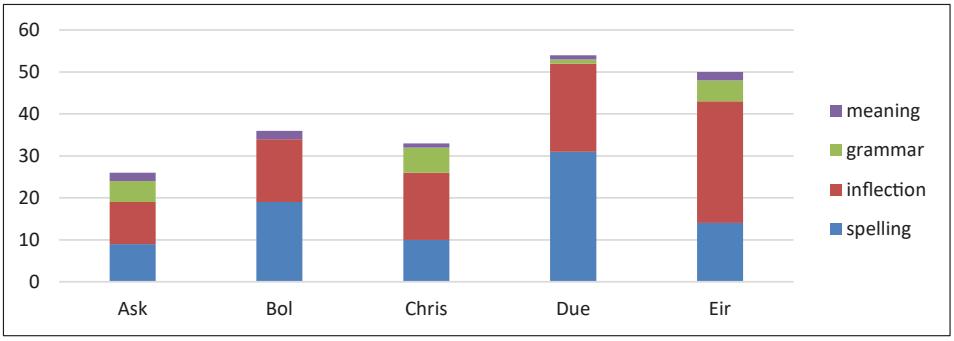


Fig. 2: Information type in queries; query intention

Queries

Almost all queries consisted of only one word. All participants other than Eir looked up *bilde* (= picture) and *skriva/e* (= write) in the dictionary; otherwise, there was little overlap. Verbs and nouns represent most of the language problems (31% and 32%, respectively) the participants consulted aids for. A closer look at the forms the participants used shows that Ask, Bol, and Eir frequently used inflected forms – present tense, infinitive with "to" (*å bruke* = to use), nouns with an indefinite article (*ein person* = a person), and plural form (*virkemidlar* = means). It looks like these participants use the dictionary as if it was a search engine. Due queried other word forms than base only twice (*openbart* = obvious(ly) TF, *inneheld* = contain, CF[8]).

For 48 queries, none or no relevant results were returned. Chris had substantially fewer such queries, as shown in Figure 3, probably because this participant never used forms other than base forms and had only one typo. Typos and using a different form than the base form

[7] The nature of the task presented to the participants – to correct language errors – sets some prerequisites for their focus.

[8] CF stands for correct form.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

constituted 26 occurrences. In their study on search logs, Nygaard/Fjeld (2008) found that spelling mistakes in queries were the reason for 46% of searches where Bokmålsordboka did not return any results. As the authors point out, dictionaries nowadays should propose related words when a query does not return any results.

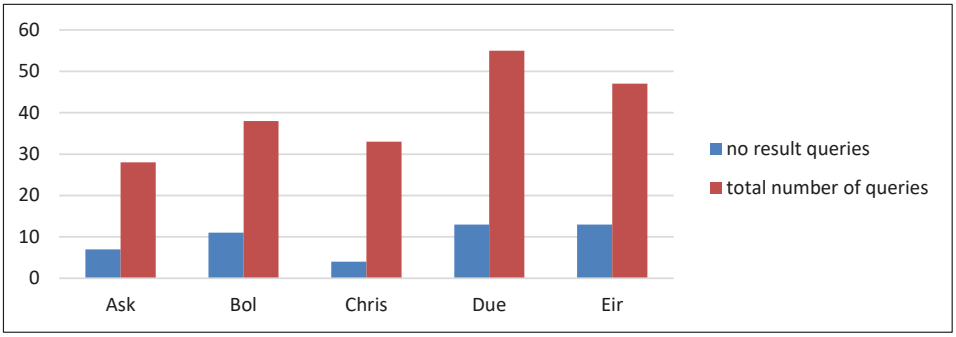


Fig. 3: Number of queries returning none or no relevant results compared to all queries carried out by the participants

The label "deviant lexical item" denotes forms that exist in Bokmål but that are not allowed in Nynorsk, for example *giftermål* (= marriage), *foran* (= in front), *hun* (= she), and *fortsatt* (= still). In these cases, users must know their Nynorsk equivalents are to be spelled *giftarmål*, *framfor/føre*, *ho*, and *framleis*, as the dictionary does not provide any hints. As Nygaard/Fjeld point out, "[t]he paradox when using a spelling dictionary is that one has to know how to spell a given word to find the actual spelling" (2008, p. 59, author's translation). Using a wildcard might have helped in such cases, but the participants never did that. Instead, they typed in different spelling(s) or proposed a synonym. Bol looked up *giftemål* (TF) in Nynorskordboka with no luck and then tried *giftermål*, which did not return any results either. Bol ended up proposing *ekteskap* (= marriage) instead. The participant did not look it up this time, probably because they did it earlier in the experiment. Typing, for example, "gift*mål" would have returned *giftarmål*. Ask, however, did not waste time querying the deviant lexical item. Instead, this participant carried out one look-up on *giftarmål* and corrected the item. When Due searched for the erroneous *openbart* (= obvious(ly), TF), the dictionary returned no results, mainly because it is not the base form of the adjective. Typing in *openbar*, although still erroneous, would have led to the sought dictionary article on *openberr*. Instead, the participant proposed the synonym *tydeleg*. Due could have used the following incremental search; typing in "openb" in the search box would prompt the dictionary to propose "openberr" at the top of the list. It seems like Due typed too fast, and the incremental search list disappeared before it was noticed. Chris paid close attention to that list when typing a query. When this participant failed to extract relevant information on "virke", they proposed the synonym "synest" (= seem). When querying this synonym, Chris noticed that the incremental search proposed the infinitive "synast" and queried that form.

One lacune was uncovered by Due – *oppsummere* = sum up. Although there is no dictionary article on this compound in Nynorskordboka, there is one in its twin dictionary Bokmålsordboka. Such cases might lead dictionary users to the conclusion that a given word is not allowed in the Nynorsk norm (Rauset 2019). This was indeed Due's conclusion. The participant then proposed the synonym *samanfatte* and carried out a look-up on it, although the only correction that really was needed regarded the spelling from *oppsumere* to *oppsummere*.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

6.3 Are online resources helpful?

By and large, consulting Nynorskordboka proved to be helpful. Of the 148 searches, 125 were categorized as having a positive effect on the issue in question. This means that 84% of all consultations either improved a language error or validated a correct form in the student text. In fact, 64 searches did not result in any changes. Many searches were carried out "just in case". In fact, Chris looked up three items (*bruke* = use, *gje/gi* = give, *vise* = show) two times each. This evidence might point to fundamental norm insecurity, as suggested by Jansson (2007), that would not be possible to detect when only scrutinizing the finished product of the corrections.

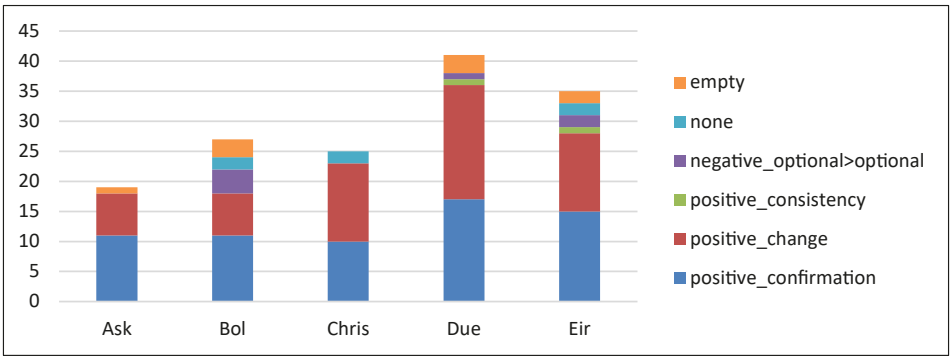


Fig. 4: Effect of consultation

Sometimes, the information in the dictionary was ignored even though it would help correct the error (annotated as "none" in Fig. 4.). Eir conducted a search on *svekka* (= weaken) to find the past participle, and even though the dictionary returned *svekt*, the participant typed in *svekka* as the correct form without commenting on their choice. While the other participants "obeyed" the dictionary, even though they stated they would have made different choices when writing themselves, Bol seemed to put personal linguistic instinct above the online resource. Bol made a change from one optional form to another twice (*brukar* to *nyttar* = use and *bilde* to *bilete* = picture). The reason Bol made these unnecessary changes was because *nyttar* and *bilete* were used in the example section. This phenomenon is not prominent in the data for the present study, but previous research has concluded that it is common for teachers to make this type of correction and thus promote a "narrower" norm with fewer allowed forms (Djupedokken 1983; Byberg 1995; Omdal 1999). This study provides evidence that this might also happen due to the word-form choices in the example section, in addition to users' inadequate dictionary skills and insufficient norm competence.

7. Closing remarks

The analysis of the students' consultation behavior shows that they almost exclusively relied on Nynorskordboka, carrying out quick searches the majority of which had a positive effect on the student text. Moreover, they tended to depend on their own linguistic instinct and knowledge rather than turning to other resources. Although not always appropriate and successful, this was a useful strategy when looking for a spelling equivalent in the dictionary because then the participants were trying to resolve the paradox of having to know a spelling of a given word to find the needed spelling. A great majority of the consulted

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

spelling and inflection problems were either improved or validated. For the other language problems (i. e., passive voice and *nokon*/*nokre*), resources other than Nynorskordboka should have been used.

Although this article is mainly concerned with consultation behavior and the effect of the consultations and not norm competency per se, several clues indicate that a higher level of metalinguistic knowledge and norm knowledge is needed. Approximately half of the words that were looked up were correct. Several errors were overlooked. The participants also showed very limited awareness of consistency requirements. The findings give evidence of norm insecurity, as suggested by Jansson (2007). Teaching dictionary skills, especially skillful querying, will speed up consultations but not improve Nynorsk writing and editing competence. Spelling and inflection are relatively easy to look up and to correct once they have been spotted. Language problems regarding consistency and syntax require a higher level of grammatical competence and prompt the use of a resource other than a dictionary.

The predominant use of Nynorskordboka stresses the importance of it meeting users' needs. One lacune was uncovered in the study (*oppsummere* = sum up). Although it is not possible to provide dictionary articles on every compound word, cases where one of the twin dictionaries has an entry and the other has not can be interpreted as the word being not within the norm. The dictionary should also be more proactive by enhancing the "Did you mean …?" function with suggestions covering deviant lexical items when no results are returned (for example *fortsatt* = still and *foran* = in front). The issue of users not being able to locate the inflection pattern is solved in the newest version of the dictionary, as it is not hidden behind a code anymore but is indicated by the label *Sjå bøying* (= show inflection). Furthermore, optional form choices in dictionary articles, although it is probably not possible to avoid, might be more normative than intended. Having in mind the authority the twin dictionaries carry among users, implementing sufficient information on norm optionality and other commonly queried issues (e. g., on passive voice) within the architecture of the dictionary should be considered.

Previous studies have stressed the importance of providing users with adequate dictionary skills. The findings of this study first and foremost indicate the necessity of improving norm knowledge and metalinguistic knowledge to make better use of aids. Teaching dictionary skills should be an integrated part of this process, but alone it may not be enough.

References

Almenningen, O./Søyland, A. (2012): Praktisk nynorsk for lærarstudentar. Oslo.

Byberg, J. (1995): Den smale veg. In: Fretland, J. O./Vikør, L. S. (eds.): Korleis bør nynorsken sjå ut? Om Norsk språkråd og normeringa av nynorsken. Oslo, pp. 37–41.

Djupedokken, S. E. (1983): Nynorsk som sidemål - Stilretting og lærarholdningar ved to vidaregåande skular i Oslo. Master thesis. Oslo.

Ericsson, K. A./Simon, H. A. (1993): Protocol analysis: Verbal reports as data. Revised edition. Cambridge, MA.

Faarlund, J. T. (2003): The Nynorsk standard language and Norwegian dialect varieties. In: Britain, D./Cheshire, J. (eds.): Social dialectology: in honour of Peter Trudgill. (= Impact: Studies in Language and Society 16). Amsterdam, pp. 311–325.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Gilquin, G./Laporte, S. (2021): The use of online writing tools by learners of English: evidence from a process corpus. In: International Journal of Lexicography 34 (4), pp. 472–492. http://doi.org/10.1093/ijl/ecab012.

Grønvik, O. et al. (2019): State-of-the-art on monolingual lexicography for Norway (Norwegian Bokmål and Nynorsk). In: Slovenščina 2.0: Empirical, Applied and Interdisciplinary Research 7 (1), pp. 39–52. http://doi.org/https://doi.org/10.4312/slo2.0.2019.1.39-52.

Helset, S. J. (2021): Norm competence among multilingual youth in Western Norway. In: Linguistic Minorities in Europe Online: A Born-Digital, Multimodal, Peer-Reviewed Online Reference. http://doi.org/https://doi.org/10.1515/lme.14813238.

Hovdenak, M./Ims, I. I. (2016): Bruk av digitale ordbøker i norsk skule. In: Nordiske Studier i Leksikografi 13, pp. 155–164.

Jansson, B. K. (2007): Ordlista – den gode hjelparen? Om rettskrivingsferidgheiter og ordlistebruk i nynorsk som sidemål. In: Nordal, A. S. (ed.): Betre nynorskundervisning. (= Skrifter frå nasjonalt senter for nynorsk i opplæringa). Høgskulen i Volda.

Karlsen, K. E./Rødningen, D. (2008): Ordboksbruk i skolen–Praksis og perspektiv Utnytting av einspråklege ordbøker i norskfaget i den vidaregåande skolen. In: LexicoNordica 15, pp. 93–114.

Laporte, S./Gilquin, G. (2018): Annotating the use of online writing resources in a video corpus of written process data in ELAN. Annotation manual version 1.1. http://hdl.handle.net/2078.1/204351.

Müller-Spitzer, C. et al. (2018): Correct hypotheses and careful reading are essential: rresults of an observational study on learners using online language resources. In: Lexikos 28, pp. 287–315.

Nygaard, L./Fjeld, R. V. (2008): Analyse av søkelogger for bedre søkemuligheter i elektroniske ordbøker. In: LexicoNordica (15), pp. 57–72.

Omdal, H. (1999): Språknormsikkerhet i bokmål og nynorsk. In: Omdal, H. (ed.): Språknormering og språkbrukar. Artiklar frå seminar ved Universitetet i Bergen. Kristiansand, pp. 181–195.

Rauset, M. (2019): Bokmålsordboka og Nynorskordboka – Einegga, toegga eller siamesiske tvillingar? In: LexicoNordica 26, pp. 155–176.

Russdal-Hamre, B. (2020): Lærarstudentar og nynorsk rettskriving. In: Juul, G. K./Helset, S. J./ Brunstad, E. (eds.): Vilkår for nynorsk mellom barn og unge. Oslo, pp. 259–282. https://doi.org/10.23865/noasp.106.

Svensén, B. (2009): A handbook of lexicography. The theory and practice of dictionary-making. Cambridge.

The Norwegian Language Council (no date-a): Rettleiing om konsekvent nynorsk. Språkrådet. http://www.sprakradet.no/sprakhjelp/Skriverad/Nynorskhjelp/Rettleiing-om-konsekvent-nynorsk/ (last access: 05-03-2020).

The Norwegian Language Council (no date-b): Nynorsk øvingsrom. Minigrammatikk. Språkrådet. http://offentlegrom.sprakradet.no/staten/minigrammatikk (last access: 16-05-2022).

The Norwegian Language Council/University of Bergen (no date): Bokmålsordboka | Nynorskordboka. https://ordbok.uib.no/ (last access: 15-11-2021).

TNS Gallup/The Norwegian Language Council (2014): Digitale ordbøker i bruk. Undersøkelse blant elever og lærere på mellom- og ungdomstrinnet og i den videregående skolen. Språkrådet. https://www.sprakradet.no/globalassets/spraka-vare/rapporter/rapport---digitale-ordboker-i-bruk.pdf.

Udir (2013): Læreplan i norsk (NOR1-05). Udir. https://www.udir.no/laring-og-trivsel/lareplanverket/finn-lareplan/lareplan/ (last access: 21-02-2022).

University of Stavanger (no date): Nordisk og lesevitenskap, master. https://www.uis.no/nb/studietilbud/nordisk-og-lesevitenskap-master (last access: 21-01-2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/ Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Wittenburg, P. et al. (2006): ELAN: A professional framework for multimodality research. In: 5th International Conference on Language Resources and Evaluation (LREC 2006), pp. 1556–1559.

Wolfer, S. et al. (2016): The effectiveness of lexicographic tools for optimising written L1-texts. In: International Journal of Lexicography 31 (1), pp. 1–28. http://doi.org/10.1093/ijl/ecw038.

Wolfer, S. et al. (2018): Combining quantitative and qualitative methods in a study on dictionary use. In: The XVIII EURALEX International Congress. Ljubljana, pp. 101–112. https://e-knjige.ff.uni-lj.si/znanstvena-zalozba/catalog/view/118/211/3000-1.pdf.

Contact information

Agnes Wigestrand Hoftun
University of Stavanger, Norway
agnes.w.hoftun@uis.no

Acknowledgements

I thank the students in the Nordic languages and literacy studies program at the University of Stavanger in Norway for choosing to participate in the study.

I thank Prof. Dr. Marte Blikstad-Balas (University of Oslo, Norway) and Dr. Stig Jarle Helset (Volda University College, Norway) for their comments on the manuscript. The assistance with and insights on the design and methodology in the early stages of the study and the comments on the manuscript provided by Dr. Sascha Wolfer (Leibniz-Institut für Deutsche Sprache, Mannheim, Germany) are greatly appreciated.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Dictionary Projects

XX EURALEX INTERNATIONAL CONGRESS
DICTIONARIES AND SOCIETY
12-16 July 2022, Mannheim, Germany

Hauke Bartels

THE LONG ROAD TO A HISTORICAL DICTIONARY OF LOWER SORBIAN

Towards a lexical information system

Abstract The Sorbian Institute has been taking preparatory steps for a historical-documentary vocabulary information system for Lower Sorbian for about 10 years. To this end, the entire extant written material (16th–21st centuries) of this strongly endangered European minority language is to be systematically evaluated. An attempt made a few years ago to organise and finance the project as a long-term scientific project was not successful in the end. Therefore, it can only be advanced step by step and via some detours. The article informs about the interim status of the project, especially with respect to the creation of a reliable database.

Keywords Lower Sorbian; historical lexicography; minority language; e-lexicography; lexical information system; text corpus; Sorbian institute; language portal

1. Introduction

After 1945, there were two attempts to secure the necessary work programme for a comprehensive historical dictionary of Sorbian through an academic project: in the 1950s and 1960s by a working group around Hans Holm Bielfeldt at the Institute for Slavonic Studies of the German Academy of Sciences (Bielfeldt 1961; Müller 1967) and again in the 2010s, with a narrower focus on Lower Sorbian, by the Sorbian Institute (Bartels 2013).[1] Both attempts were ultimately unsuccessful.

Even if it is therefore not possible to fully implement the work programme last defined at the Sorbian Institute, it has not been completely abandoned. However, the circumstances now require a very long-term and small-scale structured approach, the outcome of which is quite open. Nevertheless, the project represents the first already advanced attempt to comprehensively record and, if possible, describe the historically transmitted vocabulary of Lower Sorbian on the largest possible data basis. The current article informs about the interim status of the project and gives an outlook on the next steps.

The overarching goal of all efforts is to document Lower Sorbian as comprehensively as possible, if feasible also in conjunction with language promotion measures. The language portal dolnoserbski.de serves this purpose. Some of the language resources gathered there are mentioned below. A visual impression is given by the following screenshot:

[1] The funding application was last submitted in 2016 via the Saxon Academy of Sciences and Humanities for the 2018 research programme of the Union of the German Academies of Sciences and Humanities under the title „SorbLex Lower Sorbian. A historical-documenting vocabulary information system of the Lower Sorbian language (internet »dictionary«)“.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

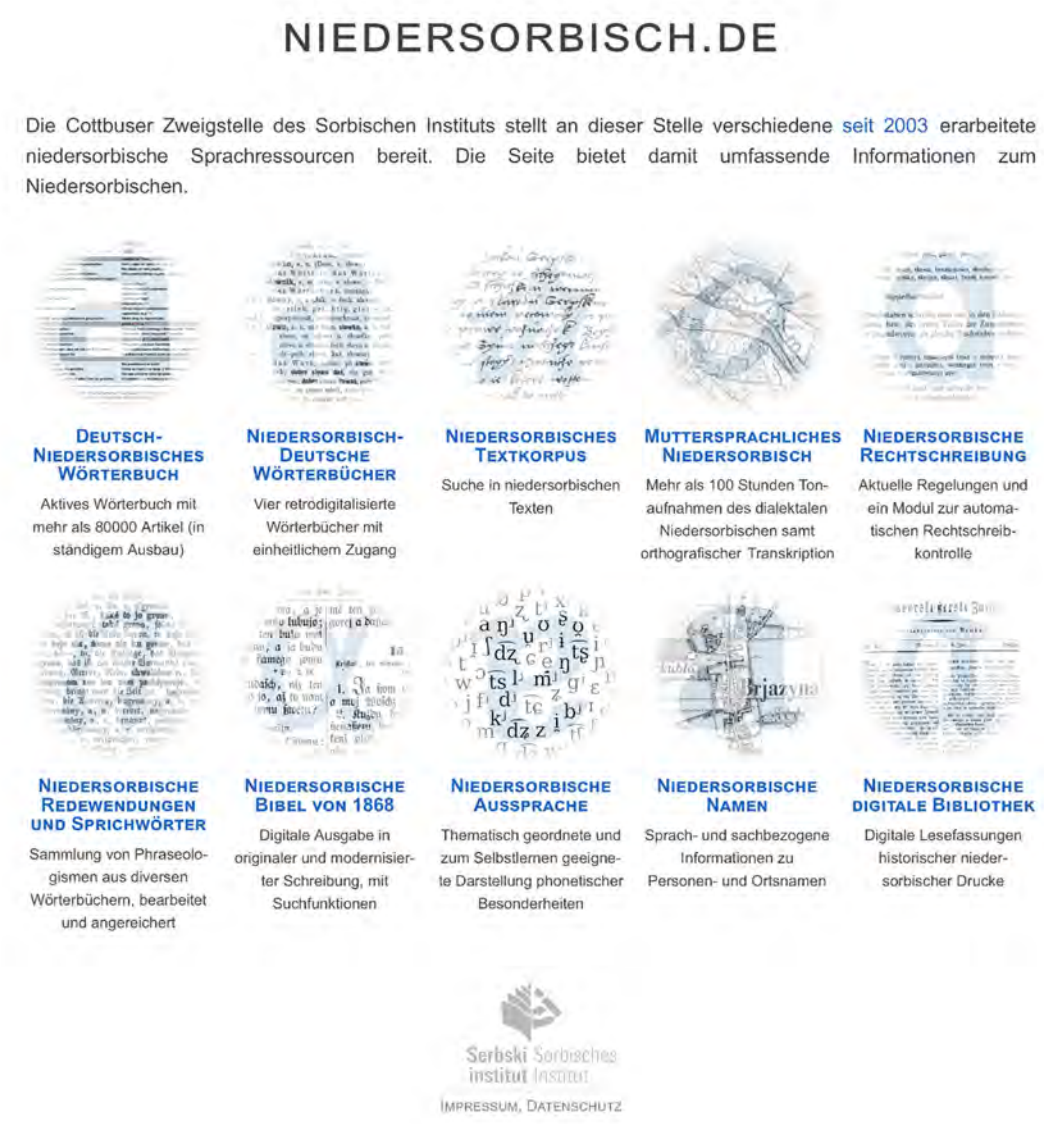


Fig. 1: The Lower Sorbian language portal (screenshot of the main page, German version)

2. Important stages on the road ...

The work programme drafted for the second academy project was divided into three phases, which – among numerous others – envisaged three main tasks:

1) Preparation of the database
2) Inventory and rough description of previously unrecorded lexis
3) Systematic re-description of the entire historical vocabulary.

The third main task has been postponed indefinitely, and the number and scope of the tasks planned for phases 1) and 2) also had to be significantly reduced. Nevertheless, important progress has been made in recent years.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

2.1 The new Lower Sorbian text corpus

For Lower Sorbian, the aim is to build up a text corpus which – as a full historical corpus – comprises the entire extant printed literature from the first half of the 16th century to the present. The construction of such a corpus began in the mid-1990s. Until around 2015, the work had to be limited to the purely quantitative expansion of the corpus. Since then, the quality of the corpus texts has also been considerably improved in various projects, which is why we now speak of the "new" Lower Sorbian text corpus. (For the history and procedure, see Bartels 2020.) By 1945, the goal of a complete corpus had already been largely achieved. Individual gaps can probably be closed in the coming years. At the same time, the digitisation and processing of post-1945 literature is progressing so that a high-quality and almost comprehensive text corpus will soon be available as a database for historical lexicography.[2] The sub-corpus of literature from the mid-20th century onwards will contain the weekly newspaper "Nowy Casnik", which was quantitatively dominant in this phase, as well as a modest proportion of journalism and fiction, and a separate section of Lower Sorbian schoolbooks. The vocabulary expected in the last mentioned sources, including large amounts of terminology, is specific in several respects, but must definitely be included in the analysis because of its great influence on several generations of pupils and thus speakers. With a view to language change in general, but especially with a view to the socio-political upheaval that took place around and after German reunification in 1990 and which also had a strong linguistic impact in various ways, it is important that the corpus also covers the decades before and after this caesura as completely as possible.

Most of the collected and digitised written material up to 1945 is freely accessible via the Lower Sorbian language portal mentioned above. Since all units of the new text corpus are of high quality and have been structurally annotated (according to the TEI-P5 guidelines), they can directly serve as the basis for a digital library.[3] At the same time, the corpus texts are also directly available for queries via the so-called "convenience search".[4] This term emphasises the benefit of the previous preparation of the corpus texts for users who would like to research a historical corpus but do not have sufficient knowledge of the grammatical and orthographical variety of forms of the text words. In this way, the text corpus, which is at the same time an important source of knowledge for all kinds of Sorbian studies, should also be opened up to non-Sorabists with as few barriers as possible.

The preparation of the corpus text for this purpose, which currently ends with a (normalising) lemmatisation and still has to do without POS tagging, for example, is above all a prerequisite for the planned lexicographic evaluation. Expansion, analysis and annotation of the corpus will continue in order to create an optimal database. An important addition to this written language corpus is the approximately 100 hours of audio recordings made between 2011 and 2016 with native speakers of Lower Sorbian of mostly very high age. These recordings were fully transcribed and are therefore also easily accessible for lexicographic evaluation.[5]

[2] The entire text corpus currently comprises about 46 million tokens, the higher-quality "new" corpus about 25 million. The total size of an all-encompassing Lower Sorbian text corpus (16th century to 2020) can be estimated at 50 to 55 million tokens.

[3] See https://www.dolnoserbski.de/biblioteka/.

[4] See https://www.dolnoserbski.de/korpus/komfort/pytanje/.

[5] See https://www.dolnoserbski.de/dobes/.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

2.2 Dictionaries as part of the data base

The second data basis for a lexical information system of Lower Sorbian are digital and retro-digitised bilingual dictionaries.[6] A first step towards an information system (more on this term in the next chapter) was the retro-digitisation of the four most important Lower Sorbian-German dictionaries (Zwahr 1847; Mucke 1911–1928; Šwjela 1961; Starosta 1999). These were modelled in XML in a uniform, fine-granular manner and made accessible in 2012 via a common user interface.[7] They represent the essential part of 150 years of Lower Sorbian lexicographic tradition and their systematic inclusion ensures, among other things, the desired transparency with regard to previous descriptions of the vocabulary.

The usability of these sources was further increased in 2014 by indexing the German-language part (more than 580,000 word forms) in addition to the usual access via Lower Sorbian lemmas in Lower Sorbian-German dictionaries. Using this "German access", not only German translation equivalents can be found, but also words in descriptions of meaning, commentaries, etc. The German-language component of the dictionary text was largely orthographically normalised and lemmatised, so that a search for *Karussell* 'carousel', for example, also records the historical spelling *Carouffel*. Such a query immediately opens up a small panorama of Lower Sorbian expressions for this traditional fairground ride, such as *karusel* and the variant *karasel*, but also the alternative terms *ẃertaw(k)a*, *kóniki*, *hobẃertawka* as well as derivatives and phrases: *kónikowy*, *karaselaŕ*, *na kónikach sejźeś*, *kóniki su pśišli*, źomy *na kóniki*. This approach is interesting not only for lexical but also for cultural-historical research, since Mucke's large dictionary in particular contains a great deal of "encyclopaedic" information.

Another component of these four retro-digitised dictionaries, which has been processed in a special way and thus made more accessible, are idioms and proverbs. These are numerous in the sources, but often difficult to find. Separate modelling and the addition of lexicographic information (for example, a literal German translation as well as explanations of meaning in Lower Sorbian and German), whose lexical components are also accessible via the search, make finding them much easier. Links lead directly from the references and the description to the sources.[8]

These four dictionaries as a lexicographic data source are supplemented by the new German-Lower Sorbian Internet dictionary (Deutsch-niedersorbisches Wörterbuch – DNW), which has been compiled for 20 years now and is constantly being expanded.[9] It currently contains more than 82,000 articles, some of them extensive, with numerous new Lower Sorbian expressions and a large number of example sentences (currently about 47,000). On the one hand, these, among others, make the DNW an active dictionary that thus makes an important contribution to language preservation. On the other hand, they can supplement the Lower Sorbian text corpus, admittedly as a very special part, since they were still predominantly formulated by native lexicographers.[10]

6 There are no monolingual dictionaries of Lower Sorbian.

7 See https://www.dolnoserbski.de/ndw/.

8 See https://www.dolnoserbski.de/nrs/.

9 See https://www.dolnoserbski.de/dnw/. See also Bartels (2010a).

10 The last native speaker lexicographer, Manfred Starosta, and another co-author of the DNW, Erwin Hannusch, have been retired for many years, but are still available as informants. Until 2008, the

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

2.3 Proper names

For Lower Sorbian proper names, a separate database has been built up in recent years, but it will be an integral part of the lexical information system. So far a good 1000 settlements or administrative units have been recorded. The site currently also contains information on more than 2600 surnames and about 120 forenames. Detailed information on the inventory, the sources evaluated, etc. can be found on the page itself.[11] Comprehensive information is provided on each of the proper names; selection and presentation vary according to the type of name. In the case of place names, in addition to more encyclopaedic information (including geo-coordinates and a short description on the affiliation to the Sorbian settlement area), there is information relevant to language comprehension (current, alternative and obsolete Lower Sorbian and German names and name forms) as well as information that promotes active use: derivatives (inhabitant, adjective) and the associated inflectional information in the form of a case-specific table. The entries for personal names contain the main form of the name (e. g. *Kóńcař*, *Měto* as a short form of *Mjertyn*), for family names additionally the name forms for women (*Kóńcarjowa*) and some other derivatives. In the case of first names, alternative and obsolete forms of the name, abbreviations and diminutives (in the above example, *Měćo*, *Mjećo*, *Mětko*) and augmentatives are also listed. Here too the adjectives formed by the respective name as well as specific inflection tables can be found. The site currently contains a total of more than 18,000 Lower Sorbian names and derived forms, most of them with complete inflection tables. For an effective search for names or name forms, more than 10,000 German names as equivalents of the Lower Sorbian ones are included as well. – There a plans to expand the proper names page in the future. A new function for conveying etymological information is currently being tested and further developed (Zschieschang 2021).

3. Lexical information system

In the academy application mentioned at the beginning, a detailed description was included of how a "historically-documenting vocabulary information system" of Lower Sorbian should be constructed conceptually and technically, what types of information it would contain, what forms of presentation it would offer, etc. The presentation of this concept for a polyfunctional and polyaccessive digital "dictionary", which most likely cannot be fully realised for the reasons mentioned, does not make sense here. Instead, we will briefly describe which aspects of a future information system already exist and which can probably be added to it in the coming years. Each of these additions brings the system a little closer to what the term "information system" promises.

The current Lower Sorbian language portal essentially comprises the individual resources already mentioned in chapter 2. On the one hand, these are part of the data basis for the development of a lexical information system, above all as a starting point for a new lexicographic description. On the other hand, they also represent a sub-component of the system, for example, as direct sources of previous lexicographic descriptions in the sense of lexicographic-historical transparency Bartels 2013,[12] p. 43).

work on the DNW was also supported by a circle of other mother-tongue informants. All but one of the members have since passed away.

[11] See https://www.dolnoserbski.de/mjenja/.

[12] The article very compactly presents an early version of the concept.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

So far, two aspects of the targeted information system are in the foreground of its gradual development: Firstly, the technical-conceptual linking of the individual resources. This should guarantee optimal availability of the information, independent of individual access routes. And secondly, the best possible solution of the so-called "finding problem" in historical lexicography (Reichmann 2012, p. 157 f.). In a similar way, however, this is also relevant for the use of historical corpora.

Both aspects are closely interwoven in many cases, which is why the steps taken so far often serve both goals at the same time:[13]

- the uniform data modelling and user interface for the Lower Sorbian-German dictionaries mentioned in chapter 2.2; in addition, the indexing of the German lexis contained there at various microstructure positions for "German access" to the dictionary information, and the additional information (and thus search options) for idioms and proverbs.
- a central Sorbian-language search on the portal's entry page, which initially includes the two most up-to-date dictionaries (Starosta 1999 and DNW).

Database and search mechanisms for the last mentioned step are designed in such a way that users can easily find the Lower Sorbian expression they are looking for (including multi-word expressions) even if they have only a very imprecise knowledge of the (historical) spelling. For this purpose, the system refers to graphically similar words. The query also recognises inflectional forms of Lower Sorbian words (such as those found in texts) and refers to their basic forms. Conversely, an additional search option also allows expressions with inflectional forms to be found when searching for the basic form. The query results appear in the form of word lists. For example, a search for *nan* ('father') produces a list of (currently) 29 single- and multi-word expressions, including such ones as *mama starego nana* ('great-grandmother'), which contain only an inflectional form of the search word. From these search results, one can then access the articles of the respective dictionaries, where further information on meaning, inflection, etc. can often be found.

Here it becomes apparent that for fully effective access to (historical) dictionary data as well as to the lexis from the historical full corpus, further data resources are necessary as part of the information system. For a strongly inflectional language such as Lower Sorbian, it is important to allocate the numerous inflectional forms to the respective basic form. For this purpose, an extensive database already exists with about 3.7 million inflectional forms for more than 90,000 basic forms (lexemes).[14] On the other hand, it is a matter of assigning the many historical spelling variants to a form representing this totality, for example in the sense of a construct lemma (Reichmann 2012, p. 160: "Konstruktlemma"). This database, which is currently being developed, will be the starting point for a corresponding reference system.

Further important steps in this direction would be the expansion of the central Sorbian search to the other Lower Sorbian-German dictionaries and a central German-language

[13] More information on these search options can be found on the portal pages themselves.

[14] At this point, it should be mentioned that Lower Sorbian, as a small or minority language, is also one of the so-called digitally under-resourced languages. This means that many resources cannot be adopted or purchased, but have to be developed specifically. This applies, for example, to an automatic spell checker, for which the database of Lower Sorbian inflectional forms was compiled, or – currently under development – a text-to-speech-function for Lower and Upper Sorbian.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

search in all these sources. Linking the (lemmatised) tokens from the text corpus (or from the texts of the digital library) with the dictionaries could also be another sensible step.

The originally planned complete lexicographical rewriting of Lower Sorbian vocabulary envisaged the creation of a comprehensive database that is uniformly described and modelled according to numerous other categories (for example, with regard to word formation components, word family affiliation, sense relations, etc.). On this basis, different variants of order or presentation should be made possible as well as an effective searchability of the entire dataset. It is still open how far this programme can be realised when we take the first steps towards a new description in the next few years.

4. Next steps

From what has been presented so far, it should have become clear that we already have a very good database due to the intensive work of the last decade. However, we will continue to work on supplementing and improving it. Some examples of ongoing and planned steps in this area have been mentioned above.

Beyond that, however, the lexicographic evaluation is to begin in 2023. This will initially involve identifying those words (including word formation and spelling variants as a basis for subsequent analysis) in the text corpus that have not yet been recorded lexicographically at all. This collection will be described for the first time. Among them there will be many loan words from German which, for purist reasons, have not been recorded in the dictionaries. The handling of these "non-Slavic" words in written Sorbian (Upper and Lower Sorbian), which in contrast to the dialects (vernacular) have often been replaced by words of Czech origin – mostly through the mediation of Upper Sorbian –, also plays a major role in the still unwritten history of Lower Sorbian.[15]

A longer-term goal is to trace individual stages in the development of the Lower Sorbian lexical system. To this end, it would be important, for example, to describe the lexical stock up to about the middle of the 19th century as it is reflected in the text corpus. Up to this time, the literature was almost exclusively religious (Bible, hymnbooks, sermons, etc.), whereas in 1848, with the first issue of the Lower Sorbian weekly newspaper "Bramborski Serbski Casnik"), other areas of lexis were also expanded or reflected in the literature. (Incidentally, even in this part of the literature, the way different authors deal with German loan words and other elements perceived as "vernacular" is already very different).

In addition to such an approach according to different periods of writing, a special treatment of culturally(-historically) significant parts of the vocabulary is also being considered (cf. e. g. Kämper 2016). It would be desirable if such a part of the upcoming lexical analysis could also be implemented. Despite all obstacles and detours, we are going our way. We just don't know yet how far we will get.

[15] There is no monograph on German loan words in Lower Sorbian as there is for Upper Sorbian (Bielfeldt 1933). Pohontsch (2002) is an important preliminary work for an overall presentation, because the replacement of German loanwords by Czech ones was in many cases carried out via Upper Sorbian. For loan words in Lower Sorbian, see also Bartels (2009, 2010b).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

References

Bartels, H. (2020): Das niedersorbische Globalkorpus als Ziel einer ganzheitlichen Konzeption zum Aufbau von Textkorpora. In: Lětopis 67 (2), pp. 3–44.

Bartels, H. (2013): Zur Konzeption eines historisch-dokumentierenden Wortschatz-Informationssystems des Niedersorbischen. Pläne zur Behebung eines drängenden Forschungsdesiderats. In: Kempgen, S./Wingender, M./Franz, N./Jakiša, M. (eds.): Deutsche Beiträge zum 15. Internationalen Slavistenkongress Minsk. München/Berlin/Washington, D.C., pp. 37–46.

Bartels, H. (2010a): The German-Lower Sorbian online dictionary. In: Dykstra, A./Schoonheim, T. (eds.): Proceedings of the XIV Euralex International Congress (Leeuwarden, 6–10 July 2010). Afûk, Ljouwert, pp. 1450–1462.

Bartels, H. (2010b): Das (diachrone) Textkorpus der niedersorbischen Schriftsprache als Grundlage für Sprachdokumentation und Sprachwandelforschung. In: Hansen, B./Grković-Major, J. (eds.): Diachronic Slavonic Syntax. Gradual Changes in Focus. München/Berlin/Wien, pp. 7–18.

Bartels, H. (2009): Loanwords in Lower Sorbian, a Slavic language of Germany. In: Haspelmath, M./Tadmor, U. (eds.): Loanwords in the world's Languages. A comparative handbook. Berlin/New York, pp. 304–329.

Bielfeldt, H. H. (1933): Die deutschen Lehnwörter im Obersorbischen. Leipzig.

Bielfeldt, H. H. (1961): Die sorabistischen Arbeiten des Instituts für Slawistik der Deutschen Akademie der Wissenschaften zu Berlin. In: Lětopis A, pp. 124–126.

Kämper, H. (2016): Kulturwissenschaftliche Orientierung in der Lexikologie. In: Jäger, L./Holly, W./Krapp, P. et al. (eds.): Language – culture – communication. An international handbook of linguistics as a cultural discipline. Berlin/Boston, pp. 737–747.

Mucke, E. (1911–15/1926, 1928): Wörterbuch der nieder-wendischen Sprache und ihrer Dialekte. Vol. 1: St. Petersburg 1911–15, Prag 1926; Vol. 2/3: Prag 1928.

Müller, K. (1967): Die Bedeutung des Sorbischen Thesaurus. In: Forschungen und Fortschritte. Berlin, pp. 283–285.

Pohontsch, A. (2002): Der Einfluss obersorbischer Lexik auf die niedersorbische Schriftsprache, Bautzen.

Reichmann, O. (2012): Historische Lexikographie. Ideen, Verwirklichungen, Reflexionen. An Beispielen des Deutschen, Niederländischen und Englischen. Berlin/Boston.

Starosta, M. (1999): Niedersorbisch-deutsches Wörterbuch. Bautzen.

Šwjela, B. (1961): Dolnoserbsko-němski słownik. Budyšyn.

Zschieschang, Chr. (2021): Onomastischer Wissenstransfer am Sorbischen Institut. Zwei neue Projekte. In Onomastica Lipsiensia 14, pp. 641–656.

Zwahr, J. G. (1847): Niederlausitz-wendisch-deutsches Handwörterbuch. (Photomechanical reprint. Bautzen 1989).

Contact information

Hauke Bartels
Sorbisches Institut
hauke.bartels@serbski-institut.de

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Acknowledgements

The progress described in this article on the "long road to a historical dictionary of Lower Sorbian" has only been possible over the years thanks to the dedicated and persistent cooperation of many people and the financial support of a whole series of projects by various institutions. My heartfelt thanks go to all staff and supporters. More information can be found on the pages of the Lower Sorbian language portal niedersorbisch.de.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Polona Gantar/Simon Krek

CREATING THE LEXICON OF MULTI-WORD EXPRESSIONS FOR SLOVENE

Methodology and structure

Abstract This paper describes a method for automatic identification of sentences in the Gigafida corpus containing multi-word expressions (MWEs) from the list of 5,242 phraseological units, which was developed on the basis of several existing open-access lexical resources for Slovene. The method is based on a definition of MWEs which includes information on two levels of corpus annotation: syntax (dependency parsing) and morphology (POS tagging), together with some additional statistical parameters. The resulting lexicon contains 12,358 sentences containing MWEs extracted from the corpus. The extracted sentences were analysed from the lexicographic point of view with the aim of establishing canonical forms of MWEs and semantic relations between them in terms of variation, synonymy, and antonymy.

Keywords Identification of multi-word expressions; multi-word expressions lexicon; canonical form of multi-word expressions; Slovene; digital dictionary database

1. Introduction

Multi-word units (MWEs) require a number of decisions to be made in relation to their linguistic and communicative properties both from a lexicographic and a natural language processing perspective. From a lexicographic point of view, the main issue concerning MWEs is understanding their type and using appropriate methods for their lexicographic description. For instance, lexicographers need to decide if all types of MWEs require explicit semantic interpretation in the form of an explanation or definition. In other words, they need to establish how distant the meaning of a phrase as a whole is from the meanings of its individual components. Another typical lexicographic issue is the placement of MWEs in the macro- and microstructure of a dictionary. Mistakenly, this question is thought to have been made obsolete by the decline of paper dictionaries. Rendered in a different manner, it remains relevant in e-dictionaries or in digital dictionary databases. In born-digital dictionaries (cf. Tavast et al. 2018; Tiberius et al. 2021; Gantar 2020) there is a trend to include different types of MWEs as stand-alone lexical units or headwords. However, other modern dictionaries tend to incorporate MWEs under their individual components which function as dictionary headwords; this raises the issue of establishing semantic links between the MWE as a whole and the senses of its individual components. Lastly, there are also issues related to the canonical form of a MWE: which of its possible variations represents the canonical form, what is the canonical word order, the number of components, relations between components, and in particular, how do we treat the numerous variations and syntactic transformations between semantically similar MWEs.

On the other hand, natural language processing is mainly concerned with two MWE-related tasks: the identification of MWEs in a running text, based on an existing MWE lexicon, and the extraction of MWEs from a text for lexicon building (Savary/Cordeiro/Ramisch 2019). For the first task, either a list of MWEs from existing lexicons or manually annotated corpora are used as the basis (Ramisch et al. 2020), and the result is a set of corpus sentences con-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

taining MWEs in all typical syntactic and semantic realisations. The second procedure refers to the identification of MWEs in corpora regardless of existing MWEs.

For our purposes, the first approach is more appropriate because it identifies both MWEs appearing in the expected or pre-recorded syntactic structures and their lexical realisations, as well as unregistered word combinations that are potentially lexicographically relevant. Our goal is to build a lexicon of automatically extracted MWEs and to integrate all the extracted data into the Digital Dictionary Database.

In this paper we first describe a method for identification of MWEs in the Gigafida corpus. We focus on the extraction of MWEs with phraseological properties that are characterised by semantic opacity and metaphoricity (hereafter, the term MWE will be used in this sense). The method is based on identifying MWEs in the corpus by using various types of information about individual MWEs, e.g. their syntactic structure, underlying POS information, lemmas, etc. In particular, we highlight the use of syntactic information. First, we describe the construction of an initial list of MWEs, where we specify the obligatory and optional constituents, the expected valency slots, and the order of the constituents. We then describe the process of extracting the MWEs from the corpus and the result: a MWE lexicon, which – in its first version – contains 5,242 MWE lexical units, together with morpho-syntactic information on their components, and 12,358 examples from the Gigafida 2.0 corpus containing corresponding MWEs.

In the second part of the paper, we describe lexicographic aspects of using this approach – the analysis of the extracted data. Our starting point is that the semantic value of MWEs needs to be considered in order for us to be able to establish the links between their numerous variants. Therefore, we will be interested in how MWEs with overlapping constituents and potentially similar meaning are related to each other, considering their variance and syntactic transformations. The lexicographic analysis also aims to identify linguistic characteristics which have a predominant influence on the meaning of MWEs and use this to establish semantic links (variation, synonymy, antonymy) that we want to include in the Digital Dictionary Database.

2. Methodology of MWE identification in the corpus

Our method for identifying MWEs in text corpora is based on a formal description of syntactic structures encoded in the XML format. The previous method we used for this purpose was based on the formal description of grammatical relations in the Word Sketches tool for Slovene (Krek/Kilgarriff 2006), i.e. the sketch grammar. The upgraded method also takes into account the level of syntactic annotation according to the dependency model, in addition to the underlying morphological level of annotation and some further statistical parameters. This method enables us to control the morphological properties of the MWE components as well as the syntactic relations between the components and provides the opportunity to work with both lexical and syntactic variations, which profoundly influence the choice of the canonical form of the MWE in the digital dictionary database.

2.1 The Initial list of MWEs

To create the initial list of MWEs, we used three lexical resources which are freely available and include MWE types that match the required properties: Slovene Lexical Database (Gan-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

tar et al. 2013), the Dictionary of Slovene Phrasemes (DSP) (Keber 2011), and a list of idioms that were manually tagged in the ssj500k 2.0 training corpus (Krek et al. 2020) based on guidelines set out in the PARSEME COST Action (Bhatia et al. 2017). The final list, which was edited according to the criteria described below, comprised 5,241 MWEs.

Typologically, we relied on the Digital Dictionary Database (DDD) model for the selection of MWEs that would represent the initial set. DDD includes several types of MWEs (Kosem/Krek/Gantar 2020): fixed phrases – semantically independent units or terminological units with a defined syntactic structure (e. g. *okrogla miza* = round table (event), *sir s plemenito plesnijo* = blue cheese); phraseological units – semantically independent units with a primary connotative, metaphorical and/or pragmatic role (e. g. *dati mir* – *give peace 'not to disturb'); collocations – statistically relevant phrases in predictable syntactic structures that have no independent semantic value as a whole (e. g. *angleški, slovenski, italijanski ... jezik* = English, Slovenian, Italian ... language); syntactic combinations – semantically transparent or non-transparent phrases with a predictable syntactic structure and sentence role, e. g. *na podlagi česa* = on the basis of something; *pod okriljem koga* = under the auspices of somebody. Among these types only phraseological units were taken into account in the creation of the first version of the lexicon. The list was manually edited by removing duplicate MWEs appearing either within a single source (for example, each MWE appears in the DSP as many times as the number of components it contains) or in more than one source. The list was re-arranged by identifying the number of components for each MWE, including optional lexical components (indicated by the bracketed part in the following example): *pasti komu v naročje (kot zrela hruška)* = *to fall in someone's lap (like a ripe pear) 'accidentally acquire something by luck', and components indicating abstract valency slots expressed by pronouns (*zlesti komu pod kožo* = to get under someone's skin 'to be able to understand or affect someone'). We have kept variants of MWEs on the list, e. g. *dati/dobiti zeleno luč* = to give/get green light, which are – as expected – syntactically and semantically related to each other, e. g. *začaran krog* = vicious circle <==> *vrteti/znajti se v začaranem krogu* = to be/to spin in a vicious circle <==> *pasti v začaran krog* = to fall into a vicious circle <==> *izvleči se iz začaranega kroga* = to get out of a vicious circle. We have removed other types of MWEs from the list, e. g. fixed phrases (črna skrinjica = black box) and collocations (*low price, low temperature*).

2.2 Syntactic structures

For automatic identification, the MWEs on the initial list were parsed and assigned with syntactic structures that form the basis for the extraction of corpus sentences. The formal description of a syntactic structure contains different types of information: (1) syntactic parse according to the dependency model, morphological information – POS + linguistic features (case, number etc.), and some additional statistical parameters. It is the dependency layer in particular which represents an upgrade from the sketch grammar model in the MWE extraction from the corpus; it is used for defining the scope (lexicalised, non-lexicalised elements) and internal structure (word forms, dependencies) of a MWE. In this section we describe the concept of a syntactic structure used in our formalism and its description in the XML format.

Each MWE is assigned with its syntactic structure or, more precisely, MWEs are grouped according to the syntactic structure they belong to. MWEs with the same (syntactic, morphological, etc.) characteristics are thus assigned with the same syntactic structure. A syn-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

tactic structure is defined by the <syntactic_structure> element, which requires three obligatory attributes: the structure ID – @id, human-readable code – @label and the structure type – @type. An example for the MWE *kaj ne da miru komu* 'something is bothering someone':

(1) <syntactic_structure type="other" label="z-l-gg-s2-z" id="122">

There are three types of syntactic structures: »single« (11 structures for each POS of the single words), »collocation« (82 structures), and »other« for all the other structures which do not belong to the first two. The syntactic structure is also attributed with a human-readable code, which contains a formalised combination of codes or tags used for POS corpus annotation, for each of the components, e. g. the syntactic structure:

(2) label="z-l-gg-s2-z"

for the mentioned MWE (*kaj ne da miru komu*) can be interpreted as a sequence of five components, with the first one being a pronoun (z-*kaj*), the second a particle (l-*ne*), the third a main verb (gg-*dati*), the fourth a noun in the genitive (s2-*miru*), and the fifth again a pronoun (z-*komu*). As already mentioned, MWEs with the same structure have the same IDs. For example, *začarani krog* (vicious circle) and *bela zastava* (white flag) both receive the same syntactic structure ID: p0-s0 (adjective-noun combination in all grammatical cases).

Tags described in the previous paragraph are used in the JOS Multext-East annotation system, which is also a required piece of information in the syntactic structure definition. This provides some flexibility regarding the use of different annotation systems, e. g. Universal Dependencies. Furthermore, individual syntactic structures are defined by three groups of information: (1) individual words or elements that form the MWE – »components«; (2) syntactic relations between the components – »dependencies«; and (3) possible restrictions and other information which is needed to extract the MWE – »definition«:

(3)
```
<system type="JOS">
<components order="fixed"></components>
<dependencies> </dependencies>
<definition></definition>
</system>
```

The order of the components (from 1–5 in our case) can be defined as »fixed« or »variable«, depending on the decision whether the extracted MWEs should follow the order defined in the syntactic structure definition or the one which is statistically prevalent in the corpus. An example of the variable structure is a combination of an adverb and a verb, where we want to follow corpus frequency with different types of adverbs: *debelo* [Adverb] *gledati* [Verb] (*to watch thickly ('to marvel, to be surprised') vs *priti* [Verb] *jutri* [Adverb] (to come tomorrow).

Each component could be of type »core« or »other«. Core components are an integral part of the MWE, while »other« components are used to include restrictions on the level of dependency parsing constituents, e. g. if we want to exclude the possibility that one of the MWE components is connected with a preposition.

(4)
```
<components order="fixed">
<component cid="1" type="core" label="z"/>
<component cid="2" type="core" label="l"/>
<component cid="3" type="core" label="gg"/>
<component cid="4" type="core" label="s2"/>
```


This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

```
<component cid="5" type="core" label="z"/>
</components>
```

Syntactic relations between the components are defined within the »dependency« element. Again, we use the Slovenian dependency system JOS for defining syntactic relations, which consists of three types of relations – those connecting elements within (all types of) phrases, those connecting sentence elements, and all others. For our purposes, the first and the second type are the most important ones. Phrasal relations include labels »dol« (attribute), »del« (part of predicate), »prir« (coordination); sentence element relations include labels »ena« (subject), »dve« (direct/indirect object), »tri«, and »štiri« (two types of adverbials). The visualisation of the above mentioned MWU (*kaj ne da miru komu*) in the Q-CAT corpus annotation tool (Brank 2021) shows that there is a subject-predicate (»ena«) connection between the pronoun *kaj*, which functions as the subject, and the verb *dati* (to give), and between two types of objects (»dve«) and the verb: the direct object *mir* ('peace') and the indirect object *komu* ('to somebody'). The negation particle *ne* ('not') is connected with the phrasal relation »del« (part of predicate).

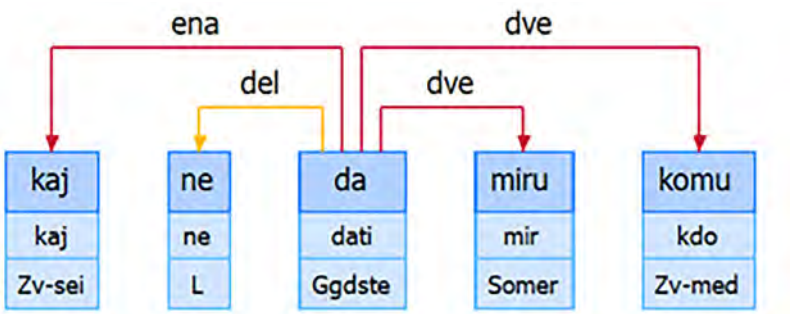


Fig. 1: A syntactically parsed MWE *kaj ne da miru komu* in the Q-Cat corpus annotation tool

The formal description of dependency relations in the XML format is the following:

(5)
```
<dependencies>
<dependency from="3" label="ena" to="1"/>
<dependency from="3" label="del" to="2"/>
<dependency from="#" label="modra" to="3"/>
<dependency from="3" label="dve" to="4"/>
<dependency from="3" label="dve" to="5"/>
</dependencies>
```

2.3 Corpus

To extract the required data, we needed a corpus annotated on various levels, based on the list of initial FEs and the definitions of syntactic structures. For this purpose, we used the Gigafida 2.1 corpus (Krek et al. 2020a), which includes additional levels of annotation, most importantly, dependency parsing annotations according to JOS (Erjavec et al. 2010, 2011) and UD (Dobrovoljc/Erjavec/Krek 2017) systems, named entities, and semantic role labelling annotations (Gantar et al. 2018).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

2.4 Extraction method

Based on the input of syntactically parsed MWEs, the script identified corpus sentences with elements included in the description of each MWE. However, we wanted to go one step further and also identify possible variants or potential new MEWs. To do this, we considered the option of filling each MWE component with any other word, as shown in Table 1.

0	**barvati (to colour)**	**kaj (something)**	**s (with)**	**črnimi (black)**	**barvami (colours)**
	x	A	C	x	x
1	slikati (to paint)	dneve (days)	s (with)	črnimi (black)	barvami (colours)
	a	a	C	C	C
2	slikati (to paint)	nevarnosti (danger)	s (with)	črnimi (black)	barvami (colours)
	a	a	C	C	C
3	barvati (to colour)	obrazke (faces)	s (with)	pisanimi (colourful)	barvami (colours)
	C	a	C	a	C
4	barvati (to colour)	jajčka (eggs)	s (with)	posebnimi (special)	barvami (colours)
	C	a	C	a	C
5	barvati (to colour)	dogajanja (events)	s (with)	črnimi (black)	odtenki (shades)
	C	a	C	C	a
6	barvati (to colour)	kozarčke (glasses)	s (with)	posebnimi (special)	barvami (colours)
	C	a	C	a	C

Table 1: List of phraseological candidates for the MWE: *barvati kaj s črnimi barvami* (paint with black colours) extracted from the Gigafida 2.1 corpus

As shown in Table 1, when extracting sentences from the corpus, we queried the corpus for all the possible realisations (1–6) for each MWE (0) in the initial dataset. We specified the MWE components which can vary (x), the fixed components (C), and the valency slots (A). In the corpus sentences (1–6), we recorded actual lexical realisations (a) in variable components (x). These lists were then used for the semantic analysis of the real occurrences of the MWEs, as attested in the written corpus of Standard Slovene, so that we could set the rules for canonical forms in the MWE lexicon and distinguish the variants and transformations of semantically related MWEs from other independent MWEs.

3. Lexicon

Data from the MWE Lexicon is integrated into the digital dictionary database as part of its data model. At the same time, it represents a specific, stand-alone resource containing 5,242 lexical units of MWE type in 12,358 examples from the Gigafida 2.0 corpus. In the Lexicon

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

(Example 6), a MWE (<lemma> within <headword>) is defined by a syntactic structure with its identification number. The components of the MWE are defined individually in a sequence (attribute=num) and represented in the <lexeme> element together with the lemma and morpho-syntactic description (attribute=msd). The <body> element records semantic information. Each MWE sense has its identification number (sense key) and a list of senses (RelatedSenseList) with definitions. Each MWE or its sense contains at least one and up to three examples from the Gigafida 2.1 corpus.

(6)
```
<entry>
<head>
<headword>
 <lemma>kaj ne da miru komu</lemma>
</headword>
 <lexicalUnit type="MWE" structure_id="122">
<component num="1">
 <lexeme lemma="kaj" msd="Zv-sei">kaj</lexeme>
</component>
<component num="2">
 <lexeme lemma="ne" msd="L">ne</lexeme>
</component>
<component num="3">
 <lexeme lemma="dati" msd="Ggdste">da</lexeme>
</component>
<component num="4">
 <lexeme lemma="mir" msd="Somer">miru</lexeme>
</component>
<component num="5">
 <lexeme lemma="kdo" msd="Zv-med">komu</lexeme>
</component>
</lexicalUnit>
</head>
<body>
<senseList>
<sense key="s.24">
<relatedSenseList>
<relatedSense senseKey="s.26"/>
</relatedSenseList>
<definitionList>
 <definition>kaj vznemirja koga; vzbuja zanimanje pri kom</definition>
</definitionList>
<exampleContainerList>
<exampleContainer>
 <corpusExample exampleId="GF9913201.308.2">Hedonistična <comp
num="1">plat</comp> vaše osebnosti <comp num="5">vam</comp> <comp
num="2">ne</comp> bo <comp num="3">dala</comp> <comp num="4">miru</
comp>, dokler ji ne boste zares prisluhnile.</corpusExample>
</exampleContainer>
</exampleContainerList>
</sense>
</senseList>
</body>
</entry>
```

The MWE Lexicon is available in the Clarin.si repository under CC BY-SA 4.0 license.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

4. Lexicographic analysis of extracted data

In section 4 we present a case study of the analysis of extracted data from a lexicographic perspective. Based on the analysis, we aim to formalise MWE descriptions in the dictionary, including the rules for canonical forms, variants, and (syntactic) transformations. All these issues are, as we will demonstrate on the example of five selected MWEs, determined by the question of their meaning. The purpose of the lexicographic analysis is thus also to identify the linguistic features which are decisive in determining the meaning of the MWEs and to incorporate these findings into a formalised description of the MWEs in the digital dictionary database and, consequently, improve the system of their extraction from the corpus.

4.1 Identification of semantic features which characterise a MWE as a lexical unit

The complexity of defining a MWE as a lexical unit, which stems from, among other things, its capacity for variation and transformation, is illustrated by an example of a MWE containing two fixed components: the verb *dati* (to give) with the variant *pustiti* (to let) and the noun *mir* (peace). There were five such MWEs in the initial list: *dati mir, dati mir komu, pustiti koga pri miru, pustiti koga na miru* and *ne dati miru* (To leave alone, to leave someone alone, to leave someone in peace, to let someone be, to not let alone / to not be able to get (someone or something) out of one's mind).

In the lexicographic analysis, the extracted data were compiled in an excel file containing the MWE ID, the status of each component with respect to the identification procedure, the actual lexical realisations of each component, and the corpus sentence containing the MWE. The table below shows the extracted data for the five selected MWEs. We list only three extracted instances for each MWE, even though there are many more extracted sentences (see last column). All of the examples were then considered for their relevance in terms of number, order, and form of components, using various editing and filtering options.

comp1	comp2	comp3	comp4	corpus sentence	numb. of sentences
dati	mir				879
dali	mir			*Mi smo skrbeli za red, oni so dali mir.* (We maintained order, and they let us be.)	
dal	mir			*Dvakrat na dan sva hodila lulat, drugače je pa pod mizo ležal in je dal mir.* (We went out to pee twice a day, but other than that he lay under the table and left me alone.)	
dati	mir	komu			446
dajte	mir	mi		*Samo ponoči mi dajte mir!* (At least leave me alone at night!)	
dam	mir	ti		*Daj mi ga, pa ti dam mir, je rekel.* (Give it to me, and I'll leave you alone.)	

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

comp1	comp2	comp3	comp4	corpus sentence	numb. of sentences
ne	dati	miru			524
ne	dali	miru		*Dokler ne bomo tega dosegli, ne bomo dali miru!* (We will not rest until we achieve this.)	
ne	dal	miru		*Nekdo vam očitno še ne bo dal miru.* (Someone is obviously not going to leave you alone just yet.)	
pustiti	koga	pri	miru		3547
pustili	ga	pri	miru	*Pustili so ga pri miru.* (They left him alone.)	
pustili	vas	pri	miru	*Tudi otroci vas ne bodo pustili pri miru, zahtevali bodo vašo pozornost in ljubezen.* (Children won't leave you alone either, they will demand your attention and love.)	
pustiti	koga	na	miru		483
pustijo	te	na	miru	Če ne delaš problemov, te pustijo na miru. (If you don't cause any problems, they leave you alone.)	
pustite	ženske	na	miru	Ženske pustite na miru! (Leave the women alone!)	

Table 2: A sample of extracted phraseological candidates for 5 MWEs with components *dati*/*pusti* and *mir*, from the Gigafida 2.1 corpus

4.1.1 Frequency and word forms

Important initial lexicographic conclusions can be reached simply by examining the frequency data: as the extracted sentences show, the MWE *pustiti koga pri miru* (1156) has the highest number of occurrences of the five, especially in combination with the (possible) variant with the preposition *na* (1155), which seems to be less frequent. We can assume that the MWE *dati mir* (1153) also incorporates all the sentences for *dati mir komu* (1154), which means that the same instances are extracted from the corpus for both MWEs. However, *dati mir* (1153) also includes realisations without the free valency slot (*komu*), which is important as the presence or absence of this slot may affect the overall semantic interpretation of the MWE. In relation to the two MWEs under consideration, the numerical data on the representation of verb forms also provided useful information, as in the examples *dajte no mir!* and *daj mi mir!* the imperative clearly stands out among the verb forms.

4.1.2 Negation

Negation is another feature that can influence the meaning of a MWE, as well as its scope and the properties of its components. In itself, negation is a problematic category, since it can be expressed syntactically in very different ways. In our case it is expressed through the negation particle *ne*, which is detached from the verb (i. e. it is not part of the verb, as in the case of *nimam* (I do not have) or *nisem* (I am not)). The affirmative-negative pair *dati mir* (1153) and *ne dati miru* (27) is interesting because, although seemingly opposite, the affirmative form actually denotes the opposite situation (that there is no peace), which is reflect-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

ed in the imperative forms: *Dajte mir!* (Leave us alone! Let us be!). On the other hand, the MWE with the negation particle *ne dati miru* (27) does not actually signify the negated action, but its affirmation (to bother, to harass). Accordingly, this MWE is not typically used in imperative verb forms.

4.1.3 Valency slots and semantic types

Valency slots included in the MWEs play an important role in determining their meaning. Both their presence and absence are important, as well as their lexical realisations, e. g. *pustiti koga pri miru* (leave someone (=human) alone) vs. *pustiti kaj pri miru* (leave something (non-human, activity) alone – 'not to be bothered with something'). This difference can be expressed by semantic types or other mechanisms which have been used in FrameNet (Fillmore /Johnson/Petruck 2003), Corpus Pattern Analysis (Hanks 2004; Hanks/Pustejovsky 2005), and others. In future upgrades to the Lexicon, we thus intend to define the actual realisations of the valency slots in terms of semantic types, based on the Slovene ontology of semantic types for nouns SLONEST (Kosem/Pori 2021).

The semantic role of valency slots and their lexical realisations is also obvious in the case of MWEs with negation elements. Despite their surface similarity, the MWE *kaj ne da miru komu* is in no sense semantically related to the MWE *ne dati miru*, nor to *kdo ne da miru komu*, since in the first case it signifies a disturbance caused by something, whereas in the second case this is primarily a human-caused annoyance. In the case of *kdo ne da miru komu*, the only exception is to be found with female and male agents (as a semantic role), where the MWE can be understood both in the sense of arousing interest and also as harassment: *Urban, moj sodelavec z oddelka s pijačami, mi ne da miru.* (I can't get Urban, my colleague in the drinks department, out of my mind. / Urban, my colleague in the drinks department, won't leave me alone.)

In the agent position represented by the pronoun *kaj* (something), one typically finds expressions such as conscience, curiosity, thought, question, etc. Some of the words in this position, depending on their frequency and semantic nuances, could suggest independent MWEs, e. g. žilica ne da miru komu (*his/her knack (for something) won't leave him alone = knack in the sense of 'ability, talent, interest'), *hudič/vrag ne da miru komu* (*devil doesn't leave him alone), which can be interpreted as 'to do something exciting, controversial (despite everything)'.

4.2 Determining MWE scope and canonical forms

The above described characteristics provided us with a starting point for defining the canonical forms of MWEs, with all the relevant information in terms of structures and individual components, as shown in the first section. Based on the lexicographic analysis, the initial list of 5 MWEs produced 13 independent MWEs. We justified their independence in terms of their semantic properties, their scope, and the number and sequence of components. At the same time, their semantic independence is demonstrated by the specific asymmetric links between the MWEs and their senses, as shown below. We first list the semantic descriptions or identified sense and then the individual MWEs identified through lexicographic analysis (new MWEs), as well as their semantic relations, which are illustrated by the corresponding sense number.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

(7) Identified senses

1	ne povzročati težav	not to cause problems
2	ne razgrajati; biti pri miru; biti »priden«	not to carry on; to be calm; to be „goody-goody“
3	prenehati nadlegovati koga	to stop harassing someone
4	v diskurzu: izraža zahtevo, opozorilo	in discourse: expressing a request, a warning
5	v diskurzu: izraža zavrnitev	in discourse: expressing refusal
6	ne se ukvarjati s čim, ne se dotikati ipd.	not to bother with something; not to touch, etc.
7	povzročati težave	to cause trouble
8	razgrajati; ne biti pri miru; ne biti »priden«	to carry on; not to be calm; not to be „goody-goody“
9	nadlegovati koga	to harass someone
10	vzbujati zanimanje, vznemirjati	to arouse interest; to disturb
11	(kljub vsemu) narediti kaj vznemirljivega, spornega	to do something exciting, controversial (despite everything)

(8)

initial MWE	**new MWE**	**semantic relation**
dati mir	dati mir	1, 2, (7, 8)*
	daj/dajte mir!	4, 1, 2
	daj/dajte no mir!	5
dati mir komu	dati mir komu	3 (9)
	daj mi mir!	4, 3
pustiti koga na/pri miru	pustiti koga na/pri miru	3
	pustiti kaj pri miru	6
	ne pustiti koga na/pri miru	9
ne dati miru	ne dati miru	7, 8 (1, 2)
	ne dati miru komu	9 (3)
	kaj ne da miru komu	10
	žilica ne da miru komu	10
	hudič/vrag ne da miru komu	11

* () – antonymy

5. Conclusions

As we have shown in this paper, MWEs can be problematic in several respects, both in terms of the way they are included and treated in the dictionary database, as well as in terms of automatic identification in the text. Their realisations within a text display significant variability since they they can adapt to a text in numerous ways, as their individual components can be interchanged with others within a phrase and take on different word forms. Additionally, other components may be interspersed between the MWE components, while some

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

MWEs may also appear as free combinations, i. e. without lexical meaning. We therefore believe that MWEs, like single-word units, should be treated as lemmas when they are included in the dictionary database; furthermore, we should also determine the extent of variation, i. e. the point at which the deviation from the canonical form violates the interdependence between the form and the meaning of the MWE, which is a prerequisite for the recognition of a MWE as a lexical unit in its own right.

In this paper we analysed automatically extracted sentences containing a specific MWE to identify the linguistic instruments for recognising their semantic independence. We highlighted the importance of free valency slots and their lexical or semantic values, which can be expressed in terms of semantic types, such as human, space, process, phenomenon, etc. Pragmatic usage, evident from the predominance of specific forms of the verb constituent, such as the imperative, as well as the presence of negation forms, also emerged as an important semantic identifier. Detailed lexicographic analysis was made possible by the process of MWE extraction on the basis of predefined syntactic structures, where the (relatively stable) MWE variants and transformation options were identified on the basis of frequency and other morpho-syntactic information about individual components and dependency relations between them. This enabled us to determine canonical forms of MWEs in the digital database and their semantic relations.

References

Bhatia, A./Bonial, C./Candito, M./Cap, F./Cordeiro, S./Foufi, V./Gantar, P./Giouli, V./Herrero, C./Iñurrieta, U./Ionescu, M./Maldonado, A./Mititelu, V./Monti, J./Nivre, J./Onofrei, M./Ow, V./ Parra Escartín, C./Sailer, M./Ramisch, C./Ramisch, R./Rizea M./Savary, A./Schneider, N./Stonayova, I./Stymne, S./Vaidya, A./Vincze, V./Walsh, A. (2017): PARSEME shared task 1.1 annotation guidelines (last updated on November 30, 2017). http://parsemefr.lif.univ-mrs.fr/parseme-st-guidelines/1.1/ (last access: 24-03-2022).

Brank, J. (2021): Q-CAT Corpus Annotation Tool 1.2. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1442 (last access: 23-03-2022).

Dobrovoljc, K./Erjavec, T./Krek, S. (2017): The Universal Dependencies Treebank for Slovenian. In: Erjavec, T./Piskorski, J./Pivovarova, L./ Šnajder, J./Steinberger, J./Yangarber, R. (eds.): Proceedings of the EACL Workshop. The 6th Workshop on Balto-Slavic Natural Language Processing, Valencia, 2017. Stroudsburg: The Association for Computational Linguistics, p. 33.

Erjavec, T./Krek, S./Fišer, D./Ledinek, N. (2011): Project JOS: linguistic annotation of Slovene. Institut Jožef Stefan, Odsek za tehnologije znanja. http://nl.ijs.si/jos/ (last access: 23-03-2018).

Erjavec, T./Krek, S./Arhar, Š./Fišer, D./Ledinek, N./Saksida, A./Sivec, B./Trebar, B. (2010): Oblikoskladenjske specifikacije JOS V1.1 2010-03-07. http://nl.ijs.si/jos/msd/html-sl/index.html (last access: 23-03-2018).

Fillmore, CH. J./Johnson, Ch. R./Petruck, M. R. L. (2003): Background to Framenet. In: International Journal of Lexicography 16 (3), pp. 235–250.

Gantar, P. (2020): Dictionary of modern Slovene: from Slovene lexical database to digital dictionary database. In: Rasprave Instituta za hrvatski jezik i jezikoslovlje 46 (2), pp. 589–602.

Gantar, P./Štrkalj Despot, K./Krek, S./Ljubešić, N. (2018): Towards semantic role labeling in Slovene and Croatian. In: Fišer, D./Pančur, A. (ed.): Proceedings of the Conference on Language Technologies & Digital Humanities, Ljubljana, September 20–21, 2018. Ljubljana, p. 93.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Gantar, P./Krek, S./Kosem, I./Šorli, M./Kocjančič, P./Grabnar, K./Yerošina, O./Zaranšek, P./Drstvenšek, N. (2013): Slovene lexical database 1.0. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1030 (last access: 23-03-2022).

Hanks, P./Pustejovsky, J. (2005): A pattern dictionary for natural language processing. In: Revue Française de linguistique appliquée 10 (2), pp. 63–82.

Hanks, P. (2004): Corpus pattern analysis. In: Williams, G./Vessier, S. (eds.): EURALEX 2004. Proceedings of the Eleventh EURALEX International Congress, Lorient, July 6–10, 2004. Lorient, p. 87.

Keber, J. (2011): Dictionary of Slovenian phrasemes. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1129 (last access: 23-03-2022).

Kosem, I./Pori, E. (2021): Slovenske ontologije semantičnih tipov: samostalniki. In: Kosem, I. (ed.): Kolokacije v slovenščini. 1. izd. Ljubljana, Znanstvena založba Filozofske fakultete, pp. 159–202.

Kosem, I./Krek, S./Gantar, P. (2020): Defining collocation for Slovenian lexical resources. In: Kosem, I./Ganatar, P. (eds.): Collocations in lexicography: existing solutions and future challenges, Slovenščina 2.0 8 (2). Ljubljana, pp. 1–27.

Krek, S./Arhar Holdt, Š./Erjavec, T./Čibej, J./Repar, A./Gantar, P./Ljubešić, N./Kosem, I./Dobrovoljc, K. (2020a): Gigafida 2.0: the reference corpus of written standard Slovene. In: Calzolari, N. (ed.): LREC 202: Twelfth International Conference on Language Resources and Evaluation, Marseille, May 11–16, 2020. European Language Resources Association (ELRA), p. 3340.

Krek, S./Erjavec, T./Dobrovoljc, K./Gantar, P./Arhar Holdt, Š./Čibej, J./Brank, J. (2020): The ssj500k training corpus for Slovene language processing. In: Fišer, D./Erjavec, T. (eds.): Jezikovne tehnologije in digitalna humanistika. Proceedings of the Conference on Language Technologies and Digital Humanities, Ljubljana, September 24–25 2020. Ljubljana: Institute of Contemporary History, p. 24.

Krek, S./Kilgarriff, A. (2006): Slovene word sketches. In: Erjavec, T./Žganec Gros, J. (eds.): Language technologies. Proceedings of the 9th International Multiconference Information Society IS 2006, Ljubljana, 9–10 October 2006. Ljubljana: Institut "Jožef Stefan", p. 62.

Ramisch, C./Savary, A./Guillaume, B./Waszczuk, J./Candito, M./Vaidya, A./ Barbu Mititelu, V./ Bhatia, A./Iñurrieta, U./ Giouli, V./Güngör, T./Jiang, M./Lichte, T./Liebeskind, Ch./Monti, J./Ramisch, R./Stymne, S./Walsh, A./Xu, H. (2020): Edition 1.2 of the PARSEME shared task on semi-supervised identification of verbal multiword expressions. In: Markantonatou, S./McCrae, J./Mitrović, J./ Tiberius, C./Ramisch, C./Vaidya, A./Osenova, P./Savary, A. (eds.), Proceedings of the Joint Workshop on Multiword Expressions and Electronic Lexicons, Barcelona (Online), December 2020. Barcelona, p. 107.

Savary, A./Cordeiro, S. R./Ramisch, C. (2019): Without lexicons, multiword expression identification will never fly: a position statement. In: Savary, A./Parra Escartín, C./Bond, F./Mitrović, J./Barbu Mititelu, V. (eds.): Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019), Florence, August 2nd, 2019. Florence, p. 79.

Tavast, A./Langemets, M./Kallas, J./Koppel, K. (2018): Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX. In: Krek, S./Čibej, J./Gorjanc, V./Kosem, I. (eds.): Lexicography in Global Contexts. Proceedings of the 18th EURALEX International Congress, Ljubljana, 17–21 July 2018. Ljubljana, p. 749.

Tiberius, C./Krek, S./Depuydt, K./Gantar, P./Kallas, J./Kosem, I./Rundell, M. (2021): Towards the ELEXIS data model: defining a common vocabulary for lexicographic resources. In: Kosem, I./Cukr, M./Jakubíček, M./Kallas, J./Krek, S./Tiberius, C. (eds.): eLex 2021. Proceedings of the eLex 2021 Conference, virtual, Brno, 5–7 July 2021. Brno, p. 56.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/ Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Contact Information

Polona Gantar
University of Ljubljana
apolonija.gantar@guest.arnes.si

Simon Krek
Jožef Stefan Institute
simon.krek@guest.arnes.si

Acknowledgements

This paper was written within the framework of the research project New Grammar of Modern Standard Slovene: resources and methods (J6-8256) and in the framework of the research programmes Slovene Language – Basic, Contrastive and Applied Research (P6-0215) and Language Resources and Technologies for the Slovene Language (P6-0411), funded by the Slovenian Research Agency.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Zoe Gavriilidou/Apostolos Garoufos

THE LEXICOGRAPHIC PROTOCOL OF Mikaela_Lex: A FREE ONLINE SCHOOL DICTIONARY OF GREEK ACCESSIBLE FOR VISUALLY-IMPAIRED SENIOR ELEMENTARY CHILDREN

Abstract The purpose of this paper is to present the lexicographic protocol and to report on the progress of compilation of Mikaela_Lex, which is a Greek, free online monolingual school dictionary for upper elementary students with visual impairments including 4,000 lemmata. The dictionary is equipped with new digital tools, such as the "Braille-system-keyboard, a "speech-to-text" tool, a "text-to-speech" tool and also a qwerty accessibility for visually non-impaired students.

Keywords Inclusive lexicography; blindness; visually impaired children; pedagogical lexicography Greek

1. Introduction

Print school dictionaries, like the print dictionary *Το πρώτο μου Λεξικό με εικόνες* (My first picture dictionary) (2005) including 4,000 lemmata, for children aged 6–9 or *Το Λεξικό μας* (Our Dictionary), including 800 lemmata (2005) for children aged 9–12, were compiled by the Greek Ministry of Education in the frame of the reform of school curricula. They are dictionaries designed to be used by school children and are adapted to their mental, linguistic, cultural, and encyclopaedic development (Antypa/Efthymiou/Mitsiaki 2006; Tarp/Gouws 2012).Their use in classroom may facilitate language learning, vocabulary acquisition, reading comprehension or writing skills in a way that increases language growth. However, their use in schools is not always successful, either because they do not address the exact needs of specific target groups of school-age children, or because pupils are not strategic dictionary users (Gavriilidou 2013; Gavriilidou/Konstantinidou 2021) and lack important reference skills that would allow them to make quick and successful searches in the dictionary (Tarp 2011; Chadjipapa et al. 2020).

Furthermore, in the case of specific target-groups, like for instance visually impaired children which are, in an inclusive perspective, educated in mainstream primary schools in Greece, such print school dictionaries are not accessible, while no other adapted lexicographic materials exist. Even though advances in technology have offered new opportunities (e. g. optical scanners, optical magnifiers, note-taking devices, etc.) to help visually impaired children to be independent at school or home, and compete with sighted people (Mulloy et al. 2014), ICT supported tools or e-learning resources are lacking in Greek schools to assist learning of this specific population or, when present, teachers lack training and support in order to implement them in classroom to empower their visually impaired students.

This leads to an urgent need for the creation of tools and resources tailored for the needs of children with visual impairments that would, however, be cost-effective, given that the costs

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

of creating such materials is often economically not justifiable due to a small number of visually impaired children speaking a lesser spoken language such as Greek.

To address the above-mentioned needs, offer equal access to printed information, and help remove barriers to learning of visually impaired children attending mainstream Greek schools, we are in process of compilation of Mikaela_Lex.

However, dictionaries are often the product of an unplanned or arbitrary way of compilation which results in an inconsistent presentation of data. Gouws/Prinsloo (2005, p. 9) maintain that the compilation and publication of any dictionary should follow a rigorous lexicographic process and that "such a process compels lexicographers to adhere to certain planning and organizational criteria". According to the authors the lexicographic process or protocol is a workflow of well-defined tasks, required for compiling a dictionary. Part of this workflow is the dictionary plan which, according to Gouws/Prinsloo (2005) includes two basic modules: The organization plan and the dictionary conceptualization plan.

The purpose of this paper is to present the lexicographic protocol designed and implemented for Mikaela_Lex compilation and also offer a detailed description of the dictionary. In section 2, we describe our dictionary's organization plan. Section 3 focuses on the dictionary's conceptualization plan phases. Finally, section 4 offers a detailed presentation of the dictionary's macro-structural characteristics and its general features.

2. Dictionary's organization plan

The organization plan of a dictionary includes its genuine purpose and its lexicographic functions. It should be noted that, its genuine purpose should address the needs of the target users (Tarp/Gouws 2012). Mikaela_Lex is a free, online, monolingual Greek school dictionary which is compiled to be used by upper elementary visually impaired children attending Greek schools. Its genuine purpose is to transfer, accessible information regarding the vocabulary included in Greek school textbooks, in order to ensure the linguistic empowerment of the specific target group. In other words, Mikaela_Lex is produced so that visually impaired children will have an accessible and easy to use instrument to assist them in achieving autonomous learning in classroom and a successful dictionary consultation procedure by reaching the goals that motivate searches within specific linguistic tasks (e. g. the need to search the meaning of a word that obscures reading comprehension, a synonym or antonym exercise in the textbook, new vocabulary acquisition, etc.). Therefore, the basic knowledge and communication orientated lexicographic functions (Tarp 2000) it fulfils are a) providing data about Greek language, and b) assist the users solve problems during text production and reception in L1.

In order to ensure successful dictionary consultation procedures by the targeted population in a user-perspective (Prinsloo/Gouws 1996), it was deemed necessary, prior to Mikaela_Lex's compilation, to launch a survey to gather information from visually impaired students and their parents, in order to trace their needs and expectations. The results partially determined the structure, content and presentation of the entries (see preparation phase below).

3. Dictionary's conceptualization plan

The dictionary conceptualization plan, on the other hand, consists of five separate, consecutive phases (Gouws/Prinsloo 2005) which are presented below:

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

I. The preparation phase

This phase was completed in December 2020, during which the dictionary's size and technical aspects were decided. In this phase, it was also decided that the dictionary basis would include all school textbooks available in Greek schools for lower and upper elementary children (see material acquisition phase below).

Size: Taking into account that due to its small size (only 800 entries) the already existing print school dictionary *Το Λεξικό μας* (Our Dictionary) (2005) left out of the scope of the dictionary a lot of frequent words included in Greek school textbooks compiled for the specific age group, and thus it was not responding adequately to the needs of the specific target group, it was decided, considering the results of a corpus-based study (see below), that the Mikaela_Lex would comprise around 4,000 lemmas when finished.

Lemmatization process: We opted for word lemmatization based on the frequency of occurrence in the lexicographic corpus specially compiled for this dictionary (see material acquisition phase below). No sublexical lemmata were included. Multilexical lemmata, on the other hand, were included in the lexicographic article of the semantic head of the multilexical item.

Technical decisions: Furthermore, it was decided that Mikaela_Lex would be accompanied by a novel application that would allow the visually impaired dictionary user to easily enter the online dictionary and search for any entry using only six buttons on the keyboard, as if using a Braille typewriter, with the help of a keyboard shortcut that converts it into a Braille basic keypad. Thus, the user is not obliged to use or learn the QWERTY keyboard.

There is an audio recording of the looked-up entry with the use of text-to-speech technology. More precisely, the application offers the option to read the word both letter by letter, so that the user is aware of what (s)he is typing and by the whole word.

It was also decided to include in the dictionary one more tool, the "speech-to-text tool" for children who do not know how to use either the QWERTY keyboard or the six buttons of Braille. The dictionary is also accessible for the general population of the school-aged children with the use of keyboard typing.

II. The material acquisition phase

Mikaela_Lex is a corpus-based user-oriented interface whose compilation was proceeded by a corpus building containing the content of 40 school textbooks of 1st to 6th Grades of Greek elementary schools. The software used for building and querying the corpus was AntConc. The corpus query program also provided vital statistical information on the corpus, e. g. the size (number of tokens), types (different words), average word length, sentence, length, etc. The headword selection was based on word frequency in a corpus of 1,119,424 tokens, given that "on the macrostructural level word-frequency counts is an extremely useful tool in the compilation of a lemma list for a new dictionary" (Gouws/Prinsloo 2005, p. 30). This phase was completed in August 2020.

III. The material preparation phase

In this phase, the material was sorted in order to omit tokens of the corpus that cannot be used in Mikaela_Lex and set the corpus in order, so that to proceed with macrostructural selection and present the lexical items to be included as entries in the dictionary. The macrostructural selection was performed in accordance with the typological crite-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

ria of Mikaela_Lex, in order to ensure the age appropriateness of data and definitions and resulted in 3,992 entries which were further compared manually with the entries of the two paper school dictionaries available for Greek pupils, "My first picture dictionary" (for ages 6–8) and "Our Dictionary" (for ages 9–11), in order to check whether important entries found in these dictionaries were also included in our dictionary.

Furthermore, the microstructural characteristics of the entries were determined in this phase: to ensure user friendliness and age-appropriate cognitive load, it was decided that the microstructure of the dictionary includes information about the Part of Speech, synonyms and antonyms, phraseology and word families. This phase was completed in October 2021.

IV. The material processing phase

This phase began in November 2021 and entails the creation of dictionary texts. Mikaela_Lex is the first school dictionary that is based on a rigorous lexicographic protocol. As a part of the dictionary conceptualisation plan of this protocol, we formulated a microstructural programme, which determined the nature and extent of the microstructure, the article structure and the way in which the different slots in the article were filled with data types.

There was also a systematic effort to provide age-appropriate noncircular definitions and supporting examples to best suit the needs of upper elementary children comparing to previously published school dictionaries which, due to the fact that they were not corpus based, they included a lot of age-inappropriate lemmas and non-pedagogical definitions.

This phase is foreseen to be completed in October 2023.

V. The publishing preparation phase

It was planned that the Mikaela_Lex will be made available as a free online lexicon at the +MorPhoSe lab of Democritus University of Thrace's webpage http://synmorphose.gr/index.php/en/ once completed. Once compiled, Mikaela_Lex will be submitted to usability testing, in order to check its effectiveness, efficacy and user satisfaction (Heid/Zimmermann 2012).

4. Microstructural characteristics and general features of Mikaela_Lex

Mikaela_Lex offers information about the morphological, phonetic and orthographic form of the lemmata (e.g. grammatical gender, part of speech) and their *semantics* (meaning, examples, synonyms, antonyms, word families). Lexicographic labels are frequently offered orally in the comment on semantics to give explicit contextual guidance to the target group users. Pragmatic labels are also used to help users relate an item in Mikaela_Lex to the world outside the dictionary.

Table 2 summarizes the basic features of Mikaela_Lex.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

1.	**Original form of publication**	Online dictionary
2.	**Completeness**	Under construction
3.	**Hypertextualization**	No hypertextualization
4.	**Multimedia**	Dictionary with text and audio
5.	**Dictionary access**	(1) **Braille system keyboard:** Entry "look up" with the "Braille-system-keyboard" tool, in order to assist the visually impaired students in case that they do not know how to use the QWERTY system. (2) **Speech-to-text tool:** Automatic look up with the "speech-to-text" tool, in order to help visually impaired students navigate better without using any of the keyboard-systems at all. (3) **Text-to-speech tool with VoiceRss:** Automatic pronunciation of all information included in the lexicographic article. (4) **QWERTY accessibility:** Accessibility for visually non-impaired students by QWERTY keyboard, in case that they do not desire to use the "speech-to-text" tool or they are not acquainted with the "Braille-system-keyboard" tool.
6.	**Gamification extension**	In order to combine learning with entertainment, a Gamification module is included for supporting and boosting vocabulary learning ("Fill-in-the-blank" game).
7.	**Automatic refresh tool**	Automatic refresh of the search box is activated after every look-up, in order to search for other words without moving hands from the keyboard.
8.	**Braille transcription tool**	Automatic transcription in Braille of each result for teaching or learning purposes.

Table 2: Overview of Mikaela_Lex's main features following Klosa (2013)

5. Conclusion

The major contribution of the paper lies in the rigorous presentation of the lexicographic protocol of Mikaela_Lex. We have also demonstrated the characteristics that the e-dictionary Mikaela_Lex offers to visually impaired students. It was shown that, in contrast to other Greek e-dictionaries or voice assistants, Mikaela_Lex contains not only much more digital characteristics and tools, but it also covers a wider range of new educational methods, such as the Gamification in word-teaching. Furthermore, the e-dictionary is equipped with new digital tools, such as the "Braille-system-keyboard" that promotes Braille's knowledge which becomes important to the world of the blind community. Finally, the dictionary provides a better linguistic processing of vocabulary in comparison with pre-existing dictionaries of primary education and, therefore, the dictionary could serve as a resource for diverse linguistic research for the creation or improvement of other e-dictionaries. All the above guarantee the accessibility and unambiguous retrieval of the information presented on both the macro- and microstructural level so the Mikaela_Lex, which will empower visually impaired students and will allow them to access autonomously to language knowledge.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

References

Antypa, J./Efthymiou, A./Mitsiaki, M. (2006): Mon premier dictionnaire illustré: la rédaction d'un dictionnaire scolaire Grec. In: Corino, E./Marello, C./Onesti, Ch. (eds.): Proceedings of the 12th EURALEX International Congress. Turin, pp. 383–393.

Chadjipapa, E./Gavriilidou, Z./Markos, A./Mylonopoulos, A. (2020): The effect of gender and educational level on dictionary use strategies adopted by upper-elementary and lower-secondary students attending Greek schools. In: International Journal of Lexicography 33 (4), pp. 443–462.

Efthymiou, A./Dimou, I./Mitsiaki, M./Antypa, I. (2005): Picture dictionary of A΄,B΄,C΄ elementary – my first picture dictionary [In Greek]. Ελληνικά Γράμματα. ΟΕΔΒ.

Gavriilidou, Z. (2013): Development and validation of the Strategy Inventory for Dictionary Use (S.I.D.U). In: International Journal of Lexicography 22 (2), pp. 135–154.

Gavriilidou, Z./Konstantinidou, E. (2021): The design of an explicit and integrated intervention program for pupils aged 10–12 with the aim to promote dictionary culture and strategies. In: Gavriilidou, Z./Mitis, L./Kiosses, S. (eds.): Proceedings of XIX Euralex International Congress, Democritus University of Thrace (virtual), 7–9 September 2021. Alexandroupolis, pp. 735–745.

Gouws, R./Prinsloo, D. (2005): Principles and practice of South African Lexicography. Stellenbosch.

Heid, U./Zimmermann, J. T. (2012): Usability testing as a tool for e-dictionary design: collocations as a case in point. In: Proceedings of the 15th Euralex International Congress. Oslo, pp. 661–671.

Kapsalis, G./Paschalis, A./Tsialos, A./Goulis, D. (2005): Dictionary for D', E', F' elementary. Our dictionary [In Greek]. ΟΕΔΒ.

Klosa, A. (2013): The lexicographical process (with special focus on online dictionaries), pp. 517–524.

Mulloy, A.M./Gevarter, C./Hopkins, M./Sutherland, K. S./Ramdoss, S. T. (2014): Assistive technology for students with visual impairments and blindness. In: Lancioni, G./Singh, N. (eds.): Assistive technologies for people with diverse abilities. (= Autism and Child Psychopathology Series). New York, pp. 113–156. https://doi.org/10.1007/978-1-4899-8029-8_5.

Prinsloo, D. J./Gouws, R. H. (1996): Formulating a new dictionary convention for the lemmatization of verbs in Northern Sotho. In: South African Journal of African Languages 16 (3), pp. 100–107.

Tarp, S. (2000): Theoretical challenges to LSP lexicography. In: Lexikos 10, pp. 189–208.

Tarp, S. (2011): Pedagogical lexicography: towards a new and strict typology corresponding to the present state-of-the-art. In: Lexicos 21 (1), pp. 217–231.

Tarp, S./Gouws, R. (2012): School dictionaries for first-language learners. In: Lexicos 22, pp. 333–351.

Contact information

Gavriilidou Zoe
Democritus University of Thrace
zoegab@otenet.gr

Garoufos Apostolos
Democritus University of Thrace
agaroufo@helit.duth.gr

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Vanessa González Ribao

FACHLEXIKOGRAFIE IN DIGITALEM ZEITALTER

Ein metalexikografisches Forschungsprojekt

Abstract This paper presents the methodology of a research project on the use of specialised German dictionaries. A mixed-methods research approach will help to answer the following main questions, concerning the lexicographic presentation of the data on the one hand and the data collection on the other hand: How do different systems of data organization and presentation affect the likelihood that users will correctly find and select the data they look up? And does the probability of success increase if users are familiar with the system? Which advantages and disadvantages do lexicographers and specialised languages experts see in using quantitative methods to extract terms? And are these methods accepted and considered reliable by the user community?

Keywords Specialised lexicography for the German language; user research; mixed methods

1. Kontextualisierung und Motivation

In Zusammenhang mit der Entwicklung des Internets zum Hauptmedium für Datentransfer und -repräsentation sind in den letzten Jahren neue metalexikografische Fragen entstanden, die nicht nur die Entwicklung von lexikografischen Nachschlagewerken (vgl. Hildenbrant/Klosa (Hg.) 2016; Klosa/Müller-Spitzer (Hg.) 2016), sondern auch ihre Benutzung (vgl. Müller-Spitzer (Hg.) 2014; Wolfer et al. 2018; Adarve 2020) betreffen, denn Sinn und Wesen der Lexikografie liegen vor allem in der nutzungsorientierten Aufbereitung von Sprach- und Sprachgebrauchswissen. Diese Ziele werden von digitalen Medien, die hierfür zahlreiche Möglichkeiten bieten, nachdrücklich unterstrichen. Trotzdem wird die Lage der Gegenwartslexikografie im digitalen Wandel von Krefeld et al. (2020, S. 5 f.) in Hinblick auf den heterogenen Digitalisierungsgrad (vgl. Lücke 2019) als disparat beschrieben. Im Besonderen scheint sich die einsprachige Fachlexikografie im deutschsprachigen Raum dieser digitalen Wende noch nicht richtig angeschlossen zu haben, denn die wenigsten Nachschlagewerke sind genuine Online-Fachwörterbücher.[1] Dies bestätigen Fuertes Olivera/Tarp (2014, S. 13 ff.) in ihrem – immer noch aktuellen – Bericht zur allgemeinen Lage der Fachlexikografie. Sie berichten, dass fachlexikografische Produkte, die die Möglichkeiten des digitalen Mediums zur Unterstützung ihrer Nutzenden sinnvoll einsetzen, eine Rarität sind (ebd., S. 16 ff.). Das deutet darauf hin, dass digitale Medien in dieser Disziplin vornehmlich als Informationsspeicher zur Wissensdokumentation dienen, aber deren Potenzial für Wissenstransfer und -repräsentation nicht optimal ausgeschöpft wird (vgl. Dräger 2020, S. 13).

Im Gegensatz zur allgemeinen deutschsprachigen Sprachlexikografie liegen aktuell keine evidenzbasierten, metalexikografischen Forschungsstudien in der einsprachigen Fachlexikografie vor, die den Weg in die digitale Zukunft ebnen und zur Veränderung des Status quo beitragen. Diesem Anliegen widmet sich das vorliegende Projekt. Zum einen steht die Eignung unterschiedlicher Wissensstrukturierung- und Datendarstellungsmodelle im Fokus der Fragestellung (vgl. Abschn. 2.1). Zum anderen interessieren in Verbindung mit der Da-

[1] Die meisten neuen Lexika werden als Digitalisierungen existierender Printauflagen verstanden, denn bei denen werden die Vorteile und Möglichkeiten von Webtechnologien (z. B. Informationsstrukturierung und interaktive oder multifunktionale Elemente) nicht genutzt.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

tengrundlage sowohl die Resonanz der Anwendung quantitativer Methoden für die Erhebung der Lemmata als auch die Akzeptanz dieser Vorgehensweise auf Seiten der Nutzenden und Expert*innen (vgl. Abschn. 2.2). Zur Beantwortung dieser Fragen werden mithilfe von kombinierten empirischen Methoden (vgl. Ivankova/Creswell 2009) der Befragung und der experimentellen Erhebung qualitative und quantitative Daten gesammelt (vgl. Abschn. 3). Ziel des vorliegenden Beitrags ist es, die Fragestellungen und die methodologischen Grundlagen des Projektes vorzustellen.

2. Theoretische Annahmen und Fragestellung

Das Medium spielt bei der Eingrenzung des Angebotes an lexikografischen Daten und ihrer Darstellung eine entscheidende Rolle; ein Umstand, der die Wichtigkeit weiterführender Forschung vor dem Hintergrund der digitalen Wende und damit eines Wechsels des Mediums von Print ins Digitale unterstreicht.

2.1 Der Einfluss des Mediums auf die Wissensstrukturierung und Darstellung der Inhalte

Die traditionelle Hegemonie der semasiologischen Perspektive in der Fachlexikografie geht Hand in Hand mit dem Umstand, dass Lexika lange Zeit ausschließlich als Printmedien erschienen und entsprechend konzipiert wurden. Vorwiegend weisen Lexika eine alphabetische Anordnung der Termini auf, denn dieses System erlaubt eine einfache, semantisch unspezifizierte Strukturierung einer großen Anzahl an Termini und einen schnellen Zugriff auf die lexikografischen Daten. Eine Einschränkung dieser Strukturierung ist, dass die Darstellung der Zusammenhänge zwischen den Termini von einem guten Verweissystem abhängig ist.

Eine alternative Datendarstellung bietet die onomasiologische Perspektive. Eine semantische oder thematische Strukturierung der Termini dient hierbei zur Orientierung im betreffenden Fachgebiet, sodass dadurch die internen Beziehungen zwischen den Termini nachvollzogen werden können.[2] Eine solche Vorgehensweise ist jedoch für eine Nachschlageressource in Printform u. a. aus Platzgründen schwer umsetzbar.

Im Gegensatz dazu ließen sich etliche format- und konzeptbedingte Einschränkungen der Wissensstrukturierung und Datendarstellung in einem digitalen Format vermeiden, dem ein Begriffssystem oder eine andere strukturierte Datengestaltung zugrunde liegt. Denn es weist folgende Vorteile auf: (i) es leistet eine Organisation von Wissen, die eine Darstellung von Begriffszusammenhängen ermöglicht[3] und (ii) es erlaubt trotzdem sowohl eine semasiologische als auch eine onomasiologische oder im besten Fall eine hybride Darstellungsform (vgl. Santos/Costa 2015) auf der Nutzungsoberfläche, sodass es (iii) u. a. verschiedene Zugangswege auf die Inhalte für Nutzende bereithält (vgl. Lang/Suchowolec 2020).

[2] Als aktuelles Beispiel dafür kann man an dieser Stelle das *Dorsch – Lexikon der Psychologie* (Wirtz (Hg.) 2021) erwähnen.

[3] Dies ermöglicht u. a. neben einem formbasierten Zugriff (bspw. Eingabefeld, Index u. a.) interaktive Tabellen wie in dem *Kleinen Wörterbuch der Verlaufsformen im Deutschen* oder unterschiedliche Visualisierungen wie in der Rubrik „wissenschaftliche Terminologie“ in *Grammis* anzubieten, die die Begriffszusammenhänge darstellen.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Aus lexikografischer Perspektive korrelieren die unterschiedlichen Datenzugangs- und Darstellungsmodelle mit den unterschiedlichen Bedürfnissen der Nutzenden bzw. mit den unterschiedlichen Funktionen des Wörterbuches: Formbasierte Ansätze unterstützen die Nutzenden insbesondere bei Leseverstehen-Aufgaben, während sich inhaltsbasierte Ansätze für Textproduktionszwecke und Wissenserweiterung besser eignen (vgl. González Ribao/Meliss 2015, S. 113 ff.). Dräger (2020, S. 14) berichtet hingegen über Nutzungstests zu einem Kollokationswörterbuch, die gezeigt haben, dass die von Fachleuten gelobte Struktur nicht zur Zielgruppe passt.

Außerdem zeigen Tiberius/Niestadt (2015, S. 33 f.) am Beispiel einer Nutzungsstudie zu einem Universalwörterbuch des Niederländischen, dass die Möglichkeiten der erweiterten Suche nicht ausgenutzt werden und dass Nutzende Schwierigkeiten im Umgang mit dem inhaltbasierten Zugriff auf die Daten haben. Darüber hinaus weisen Nutzende deutschsprachiger Sprachressourcen darauf hin, dass eine einfache Benutzung ein sehr wichtiges Merkmal für ein lexikografisches Online-Produkt ist (vgl. Adarve 2020, S. 16 ff.). Daran anknüpfend stellt Lew (2015, S. 8 f.) die Hypothese auf, dass die Vertrautheit der Nutzenden im Umgang mit den lexikografischen Designmodellen eine starke Einflussvariable sein kann. Diese Hypothese ist sehr relevant für den konkreten Fall der Fachlexikografie, denn das Auffinden des gesuchten Lemmas bzw. seines Inhalts könnte insbesondere bei inhaltsbasierten Datendarstellungen davon abhängig sein, wie kompetent die Nutzenden im Fachwissen bereits sind. Ein digitales Format erlaubt zwar einen Brückenschlag von Grundkenntnissen bis zum Expertenwissen, aber die Nutzenden können nur davon profitieren, wenn sie mit den Such- und Darstellungsoptionen umgehen können, die die Anpassung der lexikografischen Daten nach Zielgruppe und Nutzungssituation ermöglichen. Ebendaher entsteht der Bedarf, Auskunft über die lexikografischen Kompetenzen der Nutzenden unter digitalen Bedingungen (Recherchekompetenz, Bedürfnis im Umgang mit Online-Nachschlagewerken) einzuholen.

Daraus ergibt sich **Frage 1:** Wie können die Potenziale des digitalen Mediums genutzt werden, um eine benutzungsfreundliche Umgebung für sehr heterogene Gruppen von Nutzenden zu schaffen?

Wenn man berücksichtigt, dass Nutzende mit unterschiedlichen Fachkenntnissen nicht nur sehr unterschiedliche Fach- und Sprachfragen haben, sondern auch unterschiedliche Kompetenzen im Umgang mit den verschiedenen lexikografischen Systemen besitzen, stellen sich folgende Fragen:

Frage 1.1: Welche von diesen Systemen zur Datenstrukturierung und -darstellung eignen sich besser für die Beantwortung von welchen Fach- und Sprachfragen bzw. für welche Benutzungssituationen?

Frage 1.2 Im Umgang mit welchem System sind die anvisierten Nutzenden besser vertraut?

2.2 Der Einfluss des Mediums auf die lexikografische Datenbasis

Über die digitale Aufbereitung von Inhalten hinaus betrifft die digitale Wende den gesamten lexikografischen Prozess (vgl. Abschn. 1). Die Datengrundlage und Lemmaauswahl ist ein zentraler Schritt im lexikografischen Prozess, der aber in der vom Printformat geprägten Nachschlagewerktradition wenig Aufmerksamkeit geschenkt wurde (vgl. Brückner 2007, S. 173). Eine Ausschöpfung der digitalen Möglichkeiten in diesem Schritt bedeutet quantita-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

tive Methoden und statistische Verfahren zur Inventargewinnung heranzuziehen (vgl. Blessing/Dick/Heid 2015; Rösiger et al. 2015). Eine Möglichkeit, die sich in der Nachbardisziplin Terminologieforschung schon etabliert hat (vgl. Heylen/De Hertog 2015), ist die automatische Termextraktion (ATE). Bei einer ATE werden potenzielle Termini aus einer ausgewählten Textsammlung maschinell extrahiert. Die erste inhaltliche Eingrenzung beruht normalerweise auf statistischen Verfahren. Das daraus resultierende Inventar wird fachlich von Terminolog*innen und/oder Expert*innen überprüft und strukturiert. Diese Praxis beschränkt sich weder auf Fachwortschätze noch auf die Terminologie. Eine ATE wurde ebenfalls in einigen allgemeinen Wörterbuchprojekten angewendet wie bspw. in der Duden-Redaktion[4] oder für *elexiko* (vgl. Storjohann 2005). Sie stellt aber trotzdem ein Novum für die einsprachige Fachlexikografie im deutschsprachigen Raum dar, denn üblicherweise werden bei der Herstellung von Lexika zwei verschiedene Methoden zur Datengewinnung angewendet: Inventur bestehender Nachschlagewerke[5] oder Materialsammlung durch eine bzw. mehrere Expert*innen[6]. Diese Datengrundlage wird als Ausgangspunkt verwendet, von einer Expertengruppe weiterbearbeitet und möglicherweise erweitert. In Zusammenhang mit der Datengrundlage zeigen aktuelle Benutzungsbefragungen im deutschsprachigen Kontext (vgl. Adarve 2020, S. 16 ff.), dass Aktualität und Zuverlässigkeit den Nutzenden äußerst wichtig sind. Die hier vertretene Hypothese lautet: Eine Erhebung der Termini durch automatisierte Verfahren auf Basis von einschlägigen Fachtexten ergibt aktuelle und zuverlässige Daten.

Hierbei ergibt sich die **Frage 2:** Können Fachleute diese Hypothese bestätigen bzw. werden derartige Verfahren von Fachlexikograf*innen und Expert*innen als notwendig und sinnvoll für das Fach beurteilt? Damit verbunden sind folgende Teilfragen:

Frage 2.1: Erweisen sich computergesteuerte Verfahren mit fachlicher Qualitätsprüfung als eine transparente Forschungspraxis, die die Anerkennung der Nutzenden und der Forschungsgemeinschaft genießt? Mit anderen Worten: Tragen quantitative Methoden zur Glaubwürdigkeit und zum Prestige des lexikografischen Endprodukts bei?

Frage 2.2: Auf welcher Datenbasis sollen teilautomatisierte Verfahren wie ATE angewendet werden, damit die lexikografischen Inhalte auf die Bedürfnisse heterogener Zielgruppen abgestimmt werden können?

Ergänzend zu diesen leitenden Fragen stellen sich folgende übergreifende Fragen zur Nutzung von fachlexikografischen Online-Ressourcen: **Frage 3.1:** Wann und zu welchen Zwecken werden einsprachige Online-Lexika (der Sprachwissenschaft bzw. anderer Fächer) konsultiert? Und **Frage 3.2:** Welche Erwartungen und Anforderungen werden von Nutzenden und Expert*innen an ein Lexikon der nächsten Generation geknüpft?

3. Methodologie

Zur Beantwortung der obigen Fragen werden quantitative und qualitative Methoden angewendet und einem „mixed methods"-Ansatz folgend (vgl. Ivankova/Creswell 2009; Wolfer et al. 2018) systematisch miteinander verbunden. Zum einen wird eine metalexikografische

[4] Vgl. https://www.duden.de/ueber_duden/wie-kommt-ein-wort-in-den-duden (letzter Zugang: 02-08-2021).

[5] Beispiel dafür ist das *Lexikon Kommunikations- und Medienwissenschaft* (vgl. Bentele/Brosius/Jarren 2013, S. 9 f.)

[6] Beispiel dafür ist das *Wörterbuch medizinischer Fachbegriffe* (vgl. Duden 2012, S. 4).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Umfrage entwickelt, die Daten zu Nutzungsgewohnheiten und Erwartungen heterogener Zielgruppen erhebt und somit die Spannbreite möglicher Einflussvariablen auf die Nutzung von Lexika erfassen soll. Dabei sollen möglichst viele verschiedene potenzielle Gruppen innerhalb der Usercommunity abgedeckt werden (Makroperspektive, vgl. Abschn. 3.1). Zum anderen werden weitere zielgruppenspezifische Befragungen und Experimente durchgeführt, die es erlauben, bestimmte Einflussvariablen näher zu untersuchen (Mikroperspektive). Zielgruppe der Experimente sind Schüler*innen sowie Studierende mit unterschiedlichem Qualifikationsniveau (vgl. Absch. 3.2), denn die Fachkompetenz der Nutzenden kann einen Einfluss auf das Nutzungserlebnis und -verhalten haben (vgl. Abschn. 2.1). Expert*innen verschiedener Fächer der Sprachwissenschaft und ihrer Nachbardisziplinen stehen im Fokus von Interviews und moderierten Diskussionsrunden mit Fokusgruppen (Expert*innen und Fachlexikograf*innen), um ein Meinungsbild zu konkreten Aspekten des fachlexikografischen Prozesses zu erheben (vgl. Abschn. 3.3).

Die Kombination verschiedener Methoden dient dazu, unterschiedliche Blickwinkel einzunehmen, denn die Online-Umfrage erlaubt einen repräsentativen Querschnitt über das Nutzerverhalten, während durch Experimente und weitere zielgruppenspezifische Befragungen die Einflussstärke wesentlicher Variablen identifiziert und präzisiert werden können. Daher fließen einerseits die Erträge der Umfrage in die Experimente ein und anderseits fließen die Befunde aus beiden Untersuchungen dementsprechend in die Expert*innen-Interviews und Diskussionen mit Fokusgruppen ein. Auf diese Weise komplementieren sich die quantitativen und qualitativen Ergebnisse der Einzeluntersuchungen. Im Folgenden werden die geplanten evidenzbasierten Methoden detaillierter erörtert.

3.1 Makroperspektive: Übergreifende Befragung

In Verbindung mit der ersten Projektphase steht die Bearbeitung einer breit angelegten Online-Umfrage im Fragebogen-Format mit geschlossenen und offenen Fragen, die sich in zwei thematische Blöcke fassen lassen. Der erste Block dient zur Erhebung eines allgemeinen Meinungsbildes bezüglich Gewohnheiten, Bedürfnissen und Erwartungen der Nutzenden und deckt daher die Fragen 3.1 und 3.2 ab. Der zweite Block ist vor allem auf die Fragen 1 (1.1 und 1.2), 2.1 und 2.2 abgestimmt und soll Daten zu den fachlexikografischen Inhalten (Quellen und Datenbasis) und ihrer weiteren Aufbereitung (Wissensstrukturierung und Zugangswege) erheben. Die Umfrage wird in der sprachwissenschaftlichen Community in Deutschland und Europa verteilt und von Ende Juli bis Ende Oktober 2022 öffentlich zur Verfügung gestellt, damit sie eine möglichst heterogene Zielgruppe erreichen kann. Relevant für die Auswertung der Daten sind vor allem die folgenden demografischen Parameter: Altersgruppe, Affinität zum Fach, Bildungsstand, Herkunft, Muttersprache und Sprachniveau im Deutschen, da diese mit den Variablen Fach- und Sprachkompetenz korrelieren, die nicht nur Einfluss auf die angebotenen Inhalte, sondern auch auf ihre Darstellung haben können.

3.2 Mikroperspektive: Experimente mit Lernenden im Fokus

Damit die Einflussstärke der Variable „Sprach- und Fachkompetenz“ innerhalb der Zielgruppe „Lernende“ genauer untersucht werden kann, stehen folgende Gruppen von Lernenden im Fokus der Experimente: Schüler*innen deutscher Lerneinrichtungen sowie Bachelor- und Masterstudierende der Linguistik, Germanistik, DaF und Übersetzungswissenschaft

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

in Deutschland und anderen europäischen Ländern. Die Experimente sollen zum einen Aufschluss über den möglichen Einfluss der Erklärungsvariable des Umgangs mit unterschiedlichen Zugriffs- und Darstellungsmodellen des lexikografischen Inhaltsangebots geben (vgl. Frage 1.2, Experiment A). Zum anderen sollen sie Daten zu der Erklärungsvariable der Adäquatheit dieser Modelle für die Beantwortung von Fach- und Sprachfragen in verschiedenen Nutzungssituationen liefern (vgl. Frage 1.1, Experiment B). Diese Erhebungen dienen ebenfalls als Konkretisierung und Ergänzung zu den in der Online-Befragung erhobenen Daten. Die Experimente werden in Schulklassen bzw. universitären Kursen in Zusammenarbeit mit der jeweiligen Kursleitung durchgeführt und bestehen grundsätzlich aus verschiedenen Aufgaben zur Rezeption und Produktion von Fachtexten, die von Testpersonen verschiedener Qualifikationsniveaus mithilfe von unterschiedlichen lexikografischen Online-Ressourcen erledigt werden müssen. Als Hilfsmittel werden sowohl ausgewählte bestehende Lexika und fachlexikografische Ressourcen als auch Ad-hoc-Datendarstellungen angeboten. Dabei werden folgende Elemente getestet: Suchoptionen und Zugriff auf die Inhalte, Vernetzung und Visualisierung von inhaltlichen Beziehungen und Darstellung und Strukturierung von Inhalten (vgl. Abschn. 2.1). Da die Experimente unterschiedliche Zielsetzungen verfolgen, wird dementsprechend ein unterschiedlicher Kontrollgrad bei der Testung ausgeübt. Durch Experiment A soll die Affinität zu und die Vertrautheit der Studierenden mit bestimmten Systemen der Datenstrukturierung und Darstellung ermittelt werden. Deswegen dürfen die Testpersonen für die Erledigung der unterschiedlichen Aufgaben aus mehreren angebotenen Ressourcen frei auswählen. Ihre Vorgehensweise und Entscheidungen (z. B. Häufigkeit in der Auswahl bestimmter Ressourcen, Bearbeitungsdauer und Abfolge typischer Recherche- bzw. Arbeitsschritte) werden als Operationalisierung der abhängigen Variablen in einem geschlossenen Protokollformular festgehalten. In Experiment B muss jede Testperson sowohl eine kurze Aufgabe zum Leseverstehen als auch eine zur Textproduktion lösen. Die Testpersonen werden in zwei Gruppen aufgeteilt, erhalten aber jeweils die gleichen Aufgaben. Gruppe A bekommt ausgewählte formbasierte Hilfsmittel, Gruppe B bekommt ausgewählte inhaltsbasierte Hilfsmittel. Die Bewertung der Aufgabenresultate durch die entsprechenden Lehrkräfte soll erlauben, die Effizienz von unterschiedlichen lexikografischen Systemen hinsichtlich der verschiedenen lexikografischen Funktionen zu ermitteln.

3.3 Mikroperspektive: Befragungen mit Fachleuten im Fokus

In der letzten Untersuchungsphase wird der Fokus auf die Zielgruppe „Fachleute“ gesetzt. Damit wird bezweckt, ein Forum für transparente Diskussion und Meinungsaustausch zwischen Fachlexikograf*innen und Fachleuten bereitzustellen, denn diese Zielgruppe wird nicht nur als potenzielle Gruppe von Nutzenden, sondern auch als Gruppe von Mitwirkenden des lexikografischen Produkts berücksichtigt. Dafür soll zuerst ein Workshop mit Expert*innen der Bildungswissenschaft (Lehrkräfte und Pädagog*innen) und danach ein weiterer Workshop mit Expert*innen eines ausgewählten Bereichs der Sprachwissenschaft organisiert werden. In diesem Rahmen werden zwei verschiedene Befragungen durchgeführt. Durch eine moderierte Diskussion mit Fokusgruppen (Fachlexikograf*innen + Expert*innen des betreffen Fachgebietes), die einmal pro Workshop stattfinden wird, sollen Daten zur inhaltlichen Grundlage, Methodik und Datenstrukturierung (vgl. Fragen 1 und 2) unter dem konkreten Blickwinkel von den bereits erwähnten Fachgruppen erhoben werden. Als Inputmaterial dienen Ergebnisse aus den Experimenten und aus der Online-Befragung, auf deren Basis offene Fragen gestellt werden, die eine wissenschaftliche Auseinan-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

dersetzung ermöglichen und Spielraum für den Austausch bieten. Während in dieser offenen Befragungsform die Perspektive mehrerer Fokusgruppen konfrontiert wird, wird in einer strukturierten, aus spezifischen offenen Fragen bestehenden Befragungsform die Perspektive der Expert*innen in den Fokus gesetzt. Damit wird die Variable „Fachperspektive" operationalisiert. Dafür sollen pro Workshop ebenfalls mehrere Expert*innen-Interviews durchgeführt werden, die zur Ermittlung eines Meinungsbildes der Expert*innen bezüglich Erwartungen und Anforderungen an die Fachlexikografie (vgl. Frage 3.2) dienen. Die Erträge dieser Befragungen dienen ebenfalls als Kontrast und Ergänzung zu den in der Online-Befragung erhobenen Daten.

In diesen Untersuchungen sollen erste empirische und vor allem multimethodisch erhobene Daten gesammelt werden, anhand derer einerseits (i) fachlexikografische Daten nutzungsorientiert weiter be- und erarbeitet werden können und anderseits (ii) ein modellhaftes Verfahren für eine evidenzbasierte Erstellung von einsprachigen fachlexikografischen Ressourcen im digitalen Zeitalter entwickelt werden kann. Das vorliegende Vorhaben bezweckt, die Kommunikation zwischen Fachlexikograf*innen und Fachexpert*innen zu fördern und die Nutzenden aus ihrem Schattendasein herauszustellen. Ferner soll das Projekt einen Beitrag zur Transparenz in der Forschungspraxis und zur Weiterentwicklung der (digitalen) Fachlexikografie als Disziplin leisten.

Literatur

Adarve, G. (2020): Deutschschweizerische Umfrage zur Nutzung von (digitalen) Sprachressourcen – eine Ergebnisübersicht. Zürich. https://www.ds.uzh.ch/dam/jcr:737d7af9-34a0-45f3-a68f-2a740fa36889/Adarve%202020.pdf (Stand: 11.3.2022).

Bentele, G./Brosius, H./Jarren, O. (Hg.) (2013): Lexikon Kommunikations- und Medienwissenschaft. Wiesbaden.

Blessing, A./Dick, M./Heid, U. (2015): Automatische Verfahren zur Bewertung der Relevanz von Dokumenten für geisteswissenschaftliche Forschungsfragen. In: 2. Jahrestagung des Verbands Digital Humanities im deutschsprachigen Raum. https://dblp.org/db/conf/dhd/dhd2015.html (Stand: 11.3.2022).

Brückner, D. (2007): Zur Lemmaauswahl im Klassikerwörterbuch. In: Lexicographica 22, S. 1731–86.

Dräger, M. (2020): Auffindbarkeit, Sichtbarkeit, Usability und Zukunft von digitalen Sprachressourcen. In: Networx. 90. http://www.mediensprache.net/networx/networx-90.pdf (Stand: 17.8.2021).

Duden (2012): Duden – Wörterbuch medizinischer Fachbegriffe. Mannheim/Zürich.

Fuertes Olivera, P. A./Tarp, S. (2014): Theory and practice of specialised online dictionaries. (= Lexicographica. Series Maior 146). Berlin/Boston.

González Ribao, V./Meliss, M. (2015): Vorschläge zur Ausarbeitung eines onomasiologisch-konzeptuell orientierten Produktionswörterbuches im zweisprachigen Lernerkontext: Deutsch-Spanisch. In: Calañas Continente, J. A./Robles i Sabater, F. (Hg.): Die Wörterbücher des Deutschen: Entwicklungen und neue Perspektiven. Frankfurt a. M., S. 109–136.

Grammis (2022): grammatisches Informationssystem. https://grammis.ids-mannheim.de/ (Stand: 13.5.2022).

Heylen, K./De Hertog, D. (2015): Automatic Term Extraction. In: Kockaert, H. J./Steurs, F. (Hg.): Handbook of terminology. Bd. 1. Amsterdam/Philadelphia, S. 203–221.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Hildenbrandt, V./Klosa, A. (Hg.) (2016): Lexikographische Prozesse bei Internetwörterbüchern. (= OPAL 1/2016). Mannheim.

Ivankova, N./Creswell, J. (2009): Mixed methods. In: Heigham, J./Croker, R. (Hg.): Qualitative research in applied linguistics: a practical introduction. Houndmill, S. 135–163.

Kleines Wörterbuch der Verlaufsformen im Deutschen (2022). https://www.owid.de/wb/progdb/start.html (Stand: 13.5.2022).

Klosa, A./Müller-Spitzer, C. (Hg.) (2016): Internetlexikografie. Ein Kompendium. Berlin/Boston.

Krefeld T./Kümmet, S./Lücke, S./Berg-Weiß, A. (2020): Lexicographia Coniuncta (LexiCon): Aufbau einer webbasierten und bibliotheksgestützten lexikographischen Umgebung, Version 2 (05.05.2020, 15:08). In: Korpus im Text, Serie A, 40112. http://www.kit.gwi.uni-muenchen.de/?p=40112&v=2 (Stand: 17.8.2021).

Lang, C./Suchowolec, K. (2020): Wissensmanagement in der Praxis: Welchen Beitrag leistet deskriptive Terminologiearbeit? In: Ahrens, B./Beaton-Thome, M./Krein-Kühle, M./Krüger, R./Link, L./Wienen, U. (Hg.): Interdependenzen und Innovationen in Translation und Fachkommunikation. Berlin, S. 17–44.

Lew, R. (2015): Opportunities and limitations of user studies. In: Tiberius, C./Müller-Spitzer, C. (Hg.): Research into dictionary use/Wörterbuchbenutzungsforschung. 5. Arbeitsbericht des wissenschaftlichen Netzwerks „Internetlexikografie". (= OPAL 2/2015). Mannheim, S. 6–16.

Lücke, S. (2019): Digitalisierung. In: Methodologie, VerbaAlpina-de 19/1. https://www.verba-alpina.gwi.uni-muenchen.de/?page_id=493&db=191&letter=D#15 (Stand: 17.8.2021).

Müller-Spitzer, C. (Hg.) (2014): Using online dictionaries. (= Lexicographica. Series Major 145). Berlin/Boston.

Rösiger, I./Schäfer, J./George, T./Tannert, S./Heid, U./Dorna, M. (2015): Extracting terms and their relations from German texts: NLP tools for the preparation of raw material for e-dictionaries. In: Kosem, I./Jakubíček, M./Kallas, J./Krek, S. (Hg.): Proceedings of eLex 2015 Herstmonceux Castle, UK, S. 486–503.

Santos, C./Costa, R. (2015): Domain specificity. In: Kockaert, H. J./Steurs, F. (Hg.): Handbook of terminology. Bd. 1. Amsterdam/Philadelphia, S. 153–179.

Storjohann, P. (2005). *elexiko* – a corpus-based monolingual German dictionary. In: Hermes, Journal of Linguistics 34, S. 55–82.

Tiberius, C./Niestadt, J. (2015): Dictionary use: a case study of the ANW Dictionary. In: Tiberius, C./Müller-Spitzer, C. (Hg.): Research into dictionary use/Wörterbuchbenutzungsforschung. 5. Arbeitsbericht des wissenschaftlichen Netzwerks „Internetlexikografie". (= OPAL 2/2015). Mannheim, S. 28–35.

Wirtz, M. A. (Hg.) (2021): Dorsch – Lexikon der Psychologie. Göttingen.

Wolfer, S./Nied Curcio, M./Silva Dias, I. M./Müller-Spitzer, C./Domínguez Vázquez, M. J. (2018): Combining quantitative and qualitative methods in a study on dictionary use. In: Čibej, J./Gorjanc, V./Kosem, I./Krek, S. (Hg.): Proceedings of the XVIII EURALEX International Congress Lexicography in Global Contexts 17-21 July 2018, Ljubljana. Ljubljana, S. 101–112.

Kontaktinformationen

Vanessa González Ribao
Postdoc-Stipendiatin der Fritz Thyssen Stiftung
vanessina_gr@hotmail.com

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Danksagung

An dieser Stelle sei der Fritz Thyssen Stiftung für Wissenschaftsforschung herzlich gedankt, die das vorliegende Forschungsprojekt im Rahmen eines Postdoc-Stipendiums finanziell unterstützt. Ebenfalls bedanke ich mich herzlich bei zwei anonymen Gutachtenden und bei Angelika Wöllstein, Franziska Kretzschmar, Christian Lang und Meike Meliss für hilfreiche Kommentare zu einer früheren Version dieses Beitrags.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Peter Meyer

LEHNWORTPORTAL DEUTSCH: A NEW ARCHITECTURE FOR RESOURCES ON LEXICAL BORROWINGS

Abstract This paper presents the *Lehnwortportal Deutsch*, a new, freely accessible publication platform for resources on German lexical borrowings in other languages, to be launched in the second half of 2022. The system will host digital-native sources as well as existing, digitized paper dictionaries on loanwords, initially for some 15 recipient languages. All resources remain accessible as individual standalone dictionaries; in addition, data on words (etyma, loanwords etc.) together with their senses and relations to each other is represented as a cross-resource network in a graph database, with careful distinction between information present in the original sources and the curated portal network data resulting from matching and merging information on, e. g., lexical units appearing in multiple dictionaries. Special tooling is available for manually creating graphs from dictionary entries during digitization and for editing and augmenting the graph database. The user interface allows users to browse individual dictionaries, navigate through the underlying graph and 'click together' complex queries on borrowing constellations in the graph in an intuitive way. The web application will be available as open source.

Keywords Multilingual lexicography; lexical borrowings; graph database

1. Introduction

This paper presents the *Lehnwortportal Deutsch*, a freely accessible publication platform for resources on German lexical borrowings in other languages. The system is curated and hosted at the Leibniz Institute for the German Language (IDS) and will go live in 2022. The new system is a complete redesign of a web portal online since 2012 (Lehnwortportal 2012 sqq.). It will host digital-native sources as well as digitized paper dictionaries. The approach of the portal is in a certain sense the inversion of the direction of observation of conventional loan dictionaries. It makes it possible to comparatively examine different types of language contact situations and socio-cultural developments (change of rule, migration, technological innovation, etc.) in the light of the dynamics of lexical borrowing processes into the languages concerned. For German, language contact has historically arisen in a number of very different settings, among them long-lasting contact at population borders (e. g., German – Slovenian, German – Polish), contact through emigration (e. g., German – Romanian, German – American English), and contact through socio-cultural influence and elite exchange (e. g. German – Japanese, German – British English, German – Tok Pisin) (cf. Meyer/Engelberg 2011 for more discussion).

The platform will initially cover lexicographical data on English, French, Dutch, Swedish, Portuguese, Slovene, Czech, Slovak, Standard Polish, the Silesian and Cieszyn dialects of Polish, Hungarian, Turkish, and Tok Pisin. Dictionaries on German loanwords in the dialects of Polish and in the East Slavic languages Belarusian, Ukrainian, and Russian are currently being compiled specifically for the new platform (Meyer/Hentschel 2021; Meyer 2015, to appear). In the current project phase, there is a strong focus on European languages; the choice of resources is strongly constrained by the necessity to acquire publications rights. The selection of resources also reflects an effort to present as wide a variety of different loanword lexicographic approaches as possible. Most dictionaries record borrowings into the

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

written or standard language. In addition, however, the new portal will also include resources with a focus (i) on a specific dialect of a language (Cieszyn dialect of Polish); (ii) on the entire dialect continuum of a language (Polish); (iii) on subjective usage frequencies vis-à-vis native/standard vocabulary (for Silesian Polish); (iv) on borrowing history across multiple languages (Polish and the three East Slavic languages Ukrainian, Belarusian and Russian). Several contributions to the *Lehnwortportal Deutsch*, all related to Polish, have their origins in a long-standing cooperation of the IDS with the Institute of Slavic Studies at the Carl von Ossietzky University of Oldenburg.

2. Data Model

In an approach different from similar projects on Dutch (Uitleenwoordenbank 2015) and Italian (Osservatorio degli Italianismi nel Mondo), the new platform preserves its lexicographical sources in their original digital format and presentation.

In addition, the lexical units treated in the resources – loanwords, etyma, variants and derivatives thereof – and their relations to each other (such as borrowing, variation, derivation etc.), to their senses, and to their containing entries are modelled as a single large cross-resource network, technically a property graph (Rodriguez 2015).

The data model (Meyer/Eppinger 2019) consists of two interconnected graph 'layers' or subgraphs, one layer representing native dictionary data on a per-entry basis and the other one manually curated cross-resource information. If, for example, a lexical unit, such as a German etymon borrowed into multiple languages, appears in multiple resources, it should be represented as a single 'merged' graph node on the curated layer. In case the individual resources provide contradictory information on some property of the word, a lexicographically informed choice has to be made for the curated node, but through its connections to the various nodes on the native dictionary layer the conflicting source data is preserved as well.

Granular semantic versioning for documents (cf. SEMVER) will be enforced on various data levels (the portal, its dictionaries, their entries, and the graph nodes). Updates to the contents of the *Lehnwortportal* will periodically culminate in new releases; previous releases of the lexicographical data, including individual dictionary entries and graph data on lexical items ('infoboxes', see below), will remain accessible to the user through a unique and permanent URL. Editorial work is always done on a dedicated 'upcoming' release that is accessible and modifiable only for authorized users after login.

3. Backend and tooling

While the previous system used a relational database to emulate a graph (Meyer/Engelberg 2011), the new platform leverages the query capabilities and scalability of a native graph database with multi-model capabilities. The server backend of the application, with basic role-based administration tools, will be published as Open Source. As a basic principle, all data are stored in the database either as nodes in the graph or as generic documents, in order to make consistent and atomic updates as easy as possible. This includes configuration information on all relevant levels as well as lexicographical data such as XML documents, scan images and binary resources for multimedia content. Besides generic functionality for data and version management the application includes the following components.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

- There is a dedicated tool for easy manual creation of subgraphs, corresponding to separate entries of a lexicographical resource – typically a paper dictionary – to be digitized, and also for assigning the entry to correctly aligned parts of image scans in the case of paper dictionaries (Meyer/Eppinger 2018).
- An annotation tool assists in linking word senses to one or even multiple 'keywords' in a pre-computed large set of multilingual word embeddings such as ConceptNet Numberbatch (Speer/Chin/Havasi 2018), for the purpose of easy 'Google-style' onomasiological search through a large number (>10000) of available German keywords (Meyer/Tu 2021).
- The platform provides visual graph editing functionality that takes advantages of the possibility to set properties, such as the editing status of words or their properties, for strictly internal purposes. Generic graph database queries can algorithmically set internal 'flags' on graph nodes that must be manually checked. This makes it possible to use the graph capabilities for lexicographical bookkeeping purposes.

4. Online presentation

Lexicographical content is presented in a responsive Single Page Application that provides access to the individual dictionaries in a uniform, consistent manner, even though entry presentation and microstructure may vary considerably between dictionaries. Entries can be selected through traditional headword lists and optional filters. In addition to the dictionary-specific entry presentation, the lexical units treated in an entry, which are represented as nodes on the curated graph layer (see above), can be displayed in 'infoboxes' that detail all information available on the word in question. Infoboxes provide grammatical information and word senses, but also list related words, possibly in other resources, sorted according to their (borrowing/derivation/variation/...) relationships. Clicking on such a related word opens its infobox, such that users may navigate through the graph by simply following links. Infoboxes can also provide links to external resources, such as Google n-gram time series (Michel et al. 2011) or other online dictionaries, and enable authorized users to add comments on specific words.

End users may perform queries on the graph to create custom lists of lexical units, similar to a faceted search familiar from online retail catalogues. In the most elementary case, formulating a query just means adding some filter for words in a certain language, or in a specific dictionary, or with a certain part of speech etc. For their queries, users may use both globally available properties (such as part of speech, meaning) and resource-specific criteria. This approach is naturally extended to cover multi-word constellations specified through properties of the words themselves and their – possibly indirect or mediated – relations to each other. For advanced purposes, a visual graph query builder can be used to define arbitrarily complex configurations in the graph (Meyer 2019). This includes configurations that must be formalized via Boolean operators on paths in the graph. As a simple example, one might want to search for loanwords borrowed into language X that have not also been borrowed into language Y.[1]

[1] Queries with possibly nested Boolean operations on paths in a graph are not supported by most graph database query languages, at least not in full generality. The *Lehnwortportal* uses Gremlin, a low-level open-source graph traversal language (Rodriguez 2015) which is Turing-complete and thus fulfills this requirement.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

All search options will be available as components of a single generic low-threshold search facility. Query results can be displayed in multiple ways: as simple word lists augmented with a configurable set of additional word-related information; in tabular form (words with their properties) with filtering and sorting options; or in aggregated form as basic statistical visualizations, such as bar charts for part of speech distribution, timelines for the distribution of dates of first attestation (where available) or symbolic maps for geographical distribution of languages. For each word, its 'infobox' (see above) is accessible with a click, as well as an interactive graph visualization for the relevant subgraph representing the search result in question.

5. Conclusion and Future Directions

The new *Lehnwortportal Deutsch* explores the possibilities (Engelberg/Meyer 2015) and limitations (Meyer 2017) of a graph-based approach to editing, curating, presenting and querying heterogeneous, multilingual, highly interlinked lexicographical data. Using a multi-model graph database not only for storing cross-resource data on words and their relationships, but also for all kinds of lexicographical source data such as XML and images, and for configuration and administration information streamlines the editorial process and tooling significantly and helps to ensure data consistency. The most time-consuming part of the portal creation process is the manual digitization of existing dictionaries, which requires a translation of etymological information, often not sufficiently explicit, into the rigorous language of property graphs. Due to inconsistencies in the source dictionaries themselves and to the wealth of different etymological constellations, it is difficult or impossible to formulate hard rules for the translation process. In some cases, the scholarly character of exposition makes it hard even for experts to extract the relevant bits of information from a lengthy etymological discussion.

Besides the integration of new dictionaries and, possibly, manually collected data that fills gaps in previously digitized resources, there are plans for including new kinds of properties of lexical data, such as phonemic or morphological analyses, in the future.

The reasons for working with a property graph database instead of, say, an RDF triplestore are mainly practical. To this day, RDF is rarely used as the native format for dictionary creation, mainly for the lack of adequate tooling; property graphs are a more convenient and easily human-readable way of organizing network-like information. The graph data format is ideally suited for data integration into the Linguistic Linked Open Data (LLOD) infrastructure (cf. Declerck et al. 2020), as serializing property graphs to RDF is a straightforward procedure.

References

Declerck, T./McCrae, J./Hartung, M./Gracia, J./Chiarcos, Ch./Montiel, E./Cimiano, Ph./Revenko, A./Sauri, R./Lee, D./Racioppa, S./Nasir, J./Orlikowski, M./Lanau-Coronas, M./Fäth, Ch./Rico, M./Elahi, M.F./Khvalchik, M./Gonzalez, M./Cooney, K. (2020): Recent developments for the linguistic linked open data infrastructure. In: 12th International Language Resources and Evaluation Conference (LREC 2020). https://doi.org/10.5281/zenodo.3736949 (last access: 23-03-2022).

Engelberg, S./Meyer, P. (2015): Das Lehnwortportal Deutsch als kontaktlinguistisches Forschungsinstrument. In: Kelih, E./Fuchsbauer, J./Newerkla, S. M. (eds.): Lehnwörter im Slawischen: Empirische und crosslinguistische Perspektiven. Frankfurt a. M./Berlin/Bern/Bruxelles/New York/Oxford/Vienna, pp. 149–170.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Lehnwortportal Deutsch (2012 sqq.): lwp.ids-mannheim.de (last access: 23-03-2022).

Meyer, P. (2015): Aligning word senses and more: tools for creating interlinked resources in historical loanword lexicography. In: Kallas, J./Kosem, I./Krek, S. (eds.): Electronic lexicography in the 21st century: linking lexical data in the digital age. Proceedings of the eLex 2015 Conference, 11–13 August 2015, Herstmonceux Castle, United Kingdom. Ljubljana/Brighton, pp. 198–210.

Meyer, P. (2017): The limits of lexicographical abstraction. Some strengths and problems of the data architecture in the Lehnwortportal Deutsch. In: Heinz, M. (ed.): Osservatorio degli italianismi nel mondo. Punti di partenza e nuovi orizzonti. Atti dell'incontro OIM Firenze, Villa Medicea di Castello 20 giugno 2014. Florence, pp. 55–76.

Meyer, P. (2019): Leistungsfähige und einfache Suchen in lexikografischen Datennetzen. Ein Query Builder für lexikografische Property-Graphen. In: Sahle, P. (ed.): Digital Humanities: multimedial & multimodal. 6. Tagung des Verbands Digital Humanities im deutschsprachigen Raum e.V. (DHd 2019), Frankfurt am Main, Mainz, 25.3.2019 – 29.3.2019. Konferenzabstracts. Frankfurt a.M., pp. 312–314. https://doi.org/10.5281/zenodo.2600812 (last access: 23-03-2022).

Meyer, P. (to appear): "Lehnwortportal Deutsch", czyli "Portal zapożyczeń z języka niemieckiego" jako cyfrowe źródło informacji o kontaktach języka polskiego w dziedzinie leksyki [the Lehnwortportal Deutsch, or Portal of German Loanwords in Other Languages, as a digital information source on Polish lexical language contact]. In: Tom pokonferencyjny, VII Światowy Kongres Polonistów, Wrocław [Proceedings of the VII World Congress of Polonists, Wrocław].

Meyer, P./Engelberg, S. (2011): Ein umgekehrtes Lehnwörterbuch als Internetportal und elektronische Ressource: Lexikographische und technische Grundlagen. In: Hedeland, H./Schmidt, Th./Wörner, K. (eds.): Proceedings of GSCL Conference 2011 Hamburg. (= Arbeiten zur Mehrsprachigkeit, Serie B 96). Hamburg, pp. 169–174.

Meyer, P./Eppinger, M. (2018): fLexiCoGraph: creating and managing curated graph-based lexicographical data. In: Čibej, J./Gorjanc, V./Kosem, I./Krek, S. (eds.): Proceedings of the XVIII EURALEX International Congress. Lexicography in Global Contexts, 17–21 July, Ljubljana. Ljubljana, pp. 1017–1022.

Meyer, P./Eppinger, M. (2019): A web of loans: multilingual loanword lexicography with property graphs. In: Kosem, I./Zingano Kuhn, T. (eds.): Electronic lexicography in the 21st century (eLex 2019): Smart Lexicography. Book of abstracts. Sintra, Portugal, 1–3 October 2019. Brno, pp. 66–68.

Meyer, P./Hentschel, G. (2021): Charting a landscape of loans. An e-lexicographical project on German lexical borrowings in Polish dialects. In: Gavriilidou, Z./Mitits, L./Kiosses, S. (eds.): Proceedings of XIX EURALEX Congress: Lexicography for Inclusion, Vol. II, Komotini, Greece. Komotini, pp. 615–621.

Meyer, P./Tu, N.D.T. (2021): A word embedding approach to onomasiological search in multilingual loanword lexicography. In: Kosem, I./Cukr, M./Jakubíček, M./Kallas, J./Krek, S./Tiberius, C. (eds.): Electronic lexicography in the 21st century: post-editing lexicography. Proceedings of the eLex 2021 Conference. 5–7 July 2021 (virtual). Brno, pp. 78–91.

Michel, J.-B./Shen, Y.K./Aiden, A.P./Veres, A./Gray, M.K./Google Books Team/Pickett, J./Hoiberg, D./Clancy, D./Norvig, P./Orwant, J./Pinker, S./Nowak, M.A./Aiden, E.L. (2011): Quantitative analysis of culture using millions of digitized books. In: Science 331, pp. 176–182.

Osservatorio degli Italianismi nel Mondo. http://www.italianismi.org/ (last access: 23-03-2022).

Rodriguez, M. A. (2015): The Gremlin Graph Traversal Machine and Language. In: Cheney, J./Neumann, T. (eds.): Proceedings of the 15th Symposium on Database Programming Languages (DBPL 2015). New York, pp. 1–10.

SEMVER: Semantic Versioning for Documents 1.2.1. https://github.com/nils-tekampe/semverdoc/blob/master/semverdoc.md. (last access: 23-03-2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Speer, R./ Chin, J./Havasi, C. (2018): ConceptNet 5.5: An open multilingual graph of general knowledge. arXiv:1612.03975v2 [cs.CL], https://doi.org/10.48550/arXiv.1612.03975 (last access: 23-03-2022).

Uitleenwoordenbank (2015): – Sijs, N. van der (2015): Uitleenwoordenbank, hosted by the Instituut voor de Nederlandse Taal. uitleenwoordenbank.ivdnt.org (last access: 23-03-2022).

Contact information

Peter Meyer
Leibniz-Institut für Deutsche Sprache
meyer@ids-mannheim.de

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Iryna Ostapova/Volodymyr Shyrokov/Yevhen Kupriianov/
Mykyta Yablochkov

ETYMOLOGICAL DICTIONARY IN DIGITAL ENVIRONMENT

Abstract The digital environment represents a qualitatively new level of service for research work with linguistic information presented in dictionary form. And first of all, this applies to index systems. By dictionary indexing we mean a set of formalized rules and procedures, on the basis of which it is possible to obtain information about certain linguistic facts recorded in the dictionary. These rules are implemented in the form of user interfaces. However, one should take into account the fact that the effectiveness of automatic construction of index schemes for a digital dictionary is possible only in a sufficiently formalized environment. This article describes the method and technology of indexing the Etymological Dictionary of the Ukrainian Language (EDUL). For the language indexing of the dictionary, a special computer instrumental system (VLL – virtual lexicographic laboratory) was developed, and adapted to the structure of the EDUL and focused on the creation of indexes in automatic mode. The digital implementation of the EDUL made it possible to access the entire corpus of the dictionary text regardless of the time of publication of the corresponding volume and opened up opportunities for various digital interpretations of etymological information.

Keywords Ukrainian language; etymology; formal model; lexicographical system; etymological data base; index

1. Introduction

Etymological Dictionary of the Ukrainian Language in 7 volumes (hereinafter – EDUL) is a fundamental lexicographic work in the field of Ukrainian etymology (ESUM 1982–2012). The first volume of this dictionary was published in 1982, and the sixth in 2012. Currently, the text of the seventh volume is being formed, which is a multilingual index to the entire array of the dictionary. The text of the dictionary includes 26,165 dictionary entries, in the development of which 277 languages were used. For each of these languages, a separate index was built with the exact localizations of each word – the index unit. The total volume of the index is about 210,000 index elements (without the Ukrainian language). The dictionary is organized according to the nest principle. The total number of index elements in the nests (Ukrainian language) is about 230,000.

For the language indexing of the dictionary, a special computer instrumental system (Virtual Lexicographic Laboratory) was developed (Shyrokov 2018), and adapted to the structure of the EDUL and focused on the creation of indexes in automatic mode. The digital implementation of the EDUL made it possible to access the entire corpus of the dictionary text regardless of the publication time of the corresponding volume and opened up opportunities for various digital interpretations of etymological information.

To build the VLL, the following steps were performed:

1) Development of a formal model of the EDUL lexicographic system;
2) Preparation of a digital version of the EDUL text and identification of elements of its metalanguage marking elements of the structure of its lexicographic system;
3) Development of a database structure corresponding to the structure of the EDUL lexicographic system, taking into account the markers of its metalanguage;

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

4) Automatic conversion of the electronic text of the EDUL into a lexicographic database with a developed structure;
5) Creation of a tool package that supports user interfaces in on-line mode.

2. Etymological Dictionary of the Ukrainian Language (digital version)

2.1 Method

The digital transformation of lexicographic works requires some general theoretical framework to describe and represent the widest possible class of lexicographic objects. In our developments, we are based on the theory of lexicographic systems.

The dictionary is considered as a special type of information system – lexicographic system. It is an abstract language-information object focused on the implementation of a comprehensive information description of the lexical and grammatical structures of a particular language or set of languages (Shyrokov 2015).

The architecture of the system corresponds to the standard three-level architecture of ANSI/X3/SPARK information systems, according to which the conceptual, internal and external data levels are distinguished in the information system (Shyrokov 2014).

The types, structures, and formats of data representation, storage, and manipulation are defined internally.

At the external level, a set of procedures that allow the user to manipulate the data presented at the internal level is implemented.

The conceptual level of representation (conceptual model) is a symbolic, semantic model in which the ideas of various specialists about the subject area are integrated in an unambiguous, finite and non-contradictory way.

As a conceptual model, we use a lexicographic data model (Shyrokov 2018). Below we present it in a somewhat simplified form:

$$\{D, I_0^{\Omega}(D), V(I_0^{\Omega}(D)), \beta, \sigma[\beta], Red[V(I_0^{\Omega}(D)]\},$$

where D is the object (subject area) of modeling, in our case it is the Etymological dictionary of the Ukrainian language; $I_0(D) = \{x_i\}$ indicates the set of registry units of the dictionary (in the theory of lexicographic systems it is commonly called the set of *elementary information units*); $V(I_0(D))$ refers a set of descriptions (interpretations) of elementary information units, that is, texts of dictionary entries: (for dictionary $V(I_0(D)) = \{V(x_i)\}$ denotes matches the set of texts of dictionary entries, where $V(x_i)$ means a dictionary entry with a title word (head word) x_i; β denotes a set of structural elements abstracted as a result of the analysis of the dictionary text; σ[β] is a structure that is generated by a β certain operator σ and represents a system of meaningful relations that reflect the semantics of the subject area under consideration; restriction σ[β] on $V(x)$ generates a microstructure σ(x) dictionary entry $V(x)$; $Red[V(I_0(D)]$ is a *recursive reduction* mechanism that makes it possible to consistently identify more and more subtle details of the structure of the lexicographic system.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

2.2 Structure of the dictionary entry

The conceptual model of the dictionary is based on the analysis of the printing version of the EDUL, that is, the typographic design, organization and structure of printed texts of dictionary entries are analyzed, which are interpreted as identifiers of the corresponding elements of lexicographic structures β and σ [β].

As a basic structural element, an *etymological class* was introduced, which is a block of linear text of a dictionary entry that describes certain genetic relationships of a registered Ukrainian word. Etymological classes are distinguished according to formal criteria: a structural unit is identified as an *etymological class* if unique sign sequences used as separators can be identified in the text of a dictionary entry. The following types of etymological classes were identified for EDUL: the *class of the head word* (*HEAD*), a *class of derivatives* (*DER*), a *class of Slavic correspondences* (*SLAV*), a *language class* (*LANG*), *bibliographic class* (*BIBL*), a *reference class* (*LINK*). Note that the phrases "etymological class" and "language class" are used only as names of structural elements in the formal model of a dictionary entry, and not as linguistic terms. Each of these classes has a unique text structure, which gave us the opportunity to build a procedure for identifying the type of each etymological class in a dictionary entry by formal features.

Let's give a brief description of each class.

The head word class contains the registry (head) word itself and its specific parameters. This class is unique and is necessarily included in the dictionary entry.

The following descriptions belong to the *language class*: a) reconstructed forms of the head word or their bases at different stages of the development of the Proto-Slavic language, presented in an anti-chronological order; b) etymologically related to the registered word, the words of other Indo-European languages, starting with the phonetic and word-formation forms closest to the Proto-Slavic; c) etymologically related to the head word, the words of Semitic-Hamitic or Ural-Altaic languages; d) the etymological relation of the word has not been established, for example, "the etymology is unclear". EDUL analysis showed that there are maximum of two such classes in a dictionary entry, but we do not limit their number in our model. The analysis of the EDUL showed that there are two such classes in a dictionary entry maximum, but we do not limit their number in our model. The dictionary entry must include at least one class of this type.

The class of the head word and the *language class* make up the minimal structure of the dictionary entry. Etymological classes of other types are optional.

The class of derivatives contains words related to the registered word of the Ukrainian language, that is, the closest etymologically meanings. There can be no more than one etymological class of this type in the text of a dictionary entry.

The class of Slavic correspondences contains the correspondences of the head word from all Slavic languages in which they are recorded. There can be no more than one structural unit of this type in a dictionary entry.

Bibliographic class is a block of text containing information about scientific works that consider the etymology of the corresponding Ukrainian word or related words of other languages. The number of structural units of this type is not limited.

The *class of links* includes those text blocks that describe links with other dictionary entries.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Let's illustrate what has been said by the example of a small, but quite representative from the point of view of the structure of a dictionary entry with the head word **вуж** (*grass snake*). The text is presented in an authentic printed form:

(1) **вуж,** *вужа́к, вужа́ка, вуженя́, вужи́ха* Я, [*вужо́вник*] «змійовик» (мін.) Ж, [*гужа́к*] ЛЧерк, *уж, ужа́ка*, [*вужа́чий*] Я, *вужи́ний*, [*вужо́вий*], [*вужува́тий*] Ж; — р. бр. *уж*, п. wąż, ч. слц. užovka, вл. нл. wuž, слн. vóž; — псл. *ǫžь; — споріднене з лит. angìs «змія», прус. angis, лат. anguis, двн. unc «тс.», сірл. escung «вуж, вугор» (букв. «водяна змія»); іє. *ang^{u}(h)i-, з яким пов'язується також *вуго́р*. — Фасмер IV 150—151; Holub—Kop. 405; Machek ESJČ 673; Топоров I 86—87. — Пор. **вуго́р**[1], **вуго́р**[2].

The distribution by class can be represented as follows:

HEAD: **вуж**

DER: *вужа́к, вужа́ка, вуженя́, вужи́ха* Я, [*вужо́вник*] «змійовик» (мін.) Ж, [*гужа́к*] ЛЧерк, *уж, ужа́ка*, [*вужа́чий*] Я, *вужи́ний*, [*вужо́вий*], [*вужува́тий*] Ж

SLAV: р. бр. *уж*, п. wąż, ч. слц. užovka, вл. нл. wuž, слн. vóž

LANG: псл. *ǫžь

LANG: споріднене з лит. angìs «змія», прус, angis, лат. anguis, двн. unc «тс.», сірл. escung «вуж, вугор» (букв. «водяна змія»); іє. *ang^{u}(h)i-, з яким пов'язується також *вуго́р*

BIBL: Фасмер IV 150—151

BIBL: Holub—Kop. 405

BIBL: Machek ESJČ 673

LINK: **вуго́р**[1]

LINK: **вуго́р**[2]

The linear sequence of the text blocks of the article can be schematically represented as follows (in curly brackets are sequences of characters that play the role of separators):

HEAD{,}DER{;—}*SLAV*{;—}*LANG*{;—}*LANG*{;—}
BIBL{;}*BIBL*{;}*BIBL*{;— Пор.}*LINK*{,}*LINK*

The selection of linear text blocks corresponds to the traditional division of the text of a dictionary entry into zones.

In the text of each etymological class, the connections of the registered word with certain words of other languages are established. We will call all these words, including head words, *etymons* (we are aware of the controversy of this term and use it for our model). When analyzing the texts of etymological classes, eight parameters were identified by which etymons are described: a *marker of linguistic affiliation* (P_L), a *remark to the marker of linguistic affiliation* (P_{RL}), a *symbolic representation of etymon* (P_A), *belonging to the dialect vocabulary* (P_D), a *homonymy marker* (P_O), *interpretation* (P_S), *remark* (P_R), *bibliography* (P_B). We have listed parameters in the order in which they usually appear in the text of the corresponding etymological class. Two parameters are required: P_L (*marker of language affiliation*) and P_A (the *symbolic representation of the etymon*). These two parameters ensure the uniqueness of each etymon of the dictionary entry: etymons with the same sign form may have different linguistic affiliation (for example, for a dictionary entry **уж** р. бр. *уж*; ч. слц. užovka; вл. нл. wuž), or etymons with the same language affiliation may have different sign forms. The

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

other parameters are optional. A formal procedure is defined for each parameter, which allows you to isolate the corresponding parameter from the text for each etymological class.

The set of parameters $\{P_L, P_{RL}, P_A, P_D, P_O, P_S, P_R, P_B\}$ we will call the *etymon structure* and denote as $ETYM(e_i)$, where *ei* is the corresponding etymon; index *i* is the ordinal number of the appearance of this etymon in the text of the etymological class. We believe that the order of the parameters in the etymon structure is not essential. The ordinal number of the etymon is significant and is recorded in the database.

Not all parameters are relevant for each etymological class. The text that we identify as an etymological class uses its own subset of parameters; not every etymon is required to be described by a complete set of parameters. However, in order to achieve structural uniformity, one type of etymon structure is constructed for each class; if a certain parameter is not involved or cannot be distinguished by formal features, then an empty line of text corresponds to its meaning. The etymon structure is constructed only if it was possible to isolate P_A. Formally, we believe that at least one etymological structure corresponds to each etymological class. If a language class does not have any etymon (or it was not possible to identify it by a formal procedure), then we consider it a degenerate etymological class and an empty etymon structure corresponds to it. An example of a dictionary entry with such a language class:

(2) **[андріяк]** «опій»; — походження неясне.

Here is an example of etymon structures (only basic parameters) for the etymological classes *SLAV* and *LANG*.

(1) р. бр. уж, п. wąż, ч. слц. užovka, вл. нл. wuž, слн. vóž

(2) $ETYM(e_1)$ = {*PL* = <р.>, *PA* = <уж>}

(3) $ETYM(e_2)$ = {*PL* = <бр.>, *PA* = <уж>}

(4) $ETYM(e_3)$ = {*PL* = <п.>, *PA* = <wąż>}

(5) $ETYM(e_4)$ = {*PL* = <ч.>, *PA* = <užovka>}

(6) $ETYM(e_5)$ = {*PL* = <слц.>, *PA* = <wuž>}

(7) $ETYM(e_6)$ = {*PL* = <вл.>, *PA* = <wuž>}

(8) $ETYM(e_7)$ = {*PL* = <слц.>, *PA* = <vóž>}

(9) $ETYM(e_8)$ = {*PL* = <вл.>, *PA* = <wuž>}

(10) $ETYM(e_9)$ = {*PL* = <нл.>, *PA* = <wuž>}

(11) $ETYM(e_{10})$ = {*PL*= <нл.>, *PA* = <wuž>}

(12) псл. *ǫžь

(13) $ETYM(e_1)$ = {PL = <псл.>, PA = <*ǫžь>}

2.3 Multilingual index

In a printed dictionary, the text is organized in such a way that the head word has a font emphasis and is the first word of the dictionary entry, which ensures its structural significance by the capabilities of the printed text. In the proposed model, the structure-forming parameters of the ethymon structure are mandatory: entry into the dictionary is possible for any language and for any word in the alphabet of this language.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Each index element ind_el_k (k = 1, 2, … K; K is the number of elements in the corresponding index) has the following structure:

$ind_el_k \equiv \{e_k, lang(e_k), loc(e_k)\}$, where e_k — etymon, $lang(e_k)$ – marker of language affiliation, $loc(e_k)$ — localization of the etymon in the dictionary text.

For the printed index of the EDUL, the following form of localization of the etymon is proposed: volume number, page number (on which the etymon is printed), the head word of the corresponding dictionary entry. The numbers of volumes and pages are a tribute to tradition and binding to the printed version of the dictionary, individual volumes of which were published with a significant time interval. Before conversion to the database, the text of each article was linked to its printed original, the volume and page numbers were recorded in the corresponding database fields. Redundancy of localization parameters can be explained by the desire to combine the approaches of the digital and printed versions: the head word in the printed version is analogous to the ID of the dictionary entry; the page number is useful when the dictionary entry occupies several pages of text (this is typical for verbs). In the digital version of EDUL, the etymon is localized up to the ordinal number of the word in the etymological class string.

The task of constructing language indexes is assigned to the instrumental system. A separate index is formed for each of the 277 languages recorded in the dictionary. The index is edited in two stages: 1) first, these word structures are edited in the database at the level of a dictionary entry; 2) a file of index elements is formed on request to the database, alphabetized by head words for a given language and edited in .doc format. After editing, the file is ready for re-conversion to the database. For the publishing system, files are processed additionally by two programs: 1) the signs that are used for etymons are inventoried; 2) an alphabet is formed and index elements are ordered according to this alphabet.

There is a fragment of the indexes for the languages that are used in the dictionary entry (the names of the languages are given in the order of their appearance in the text of the dictionary entry):

Ukrainian language
вуж 1, 437 **вуж**
вужа́к 1, 437 **вуж**
вужа́ка 1, 437 **вуж**
вуженя́ 1, 437 **вуж**
вужи́ха 1, 437 **вуж**
[вужо́вник] 1, 437 **вуж**
[гужа́к] 1, 437 **вуж**
уж 1, 437 **вуж**
ужа́ка 1, 437 **вуж**
[вужа́чий] 1, 437 **вуж**
вужи́ний 1, 437 **вуж**
[вужо́вий] 1, 437 **вуж**
[вужува́тий] 1, 437 **вуж**
…

Russian language
уж 1, 437 **вуж**
Belarusian language
уж 1, 437 **вуж**
Polish language
1, 437 **вуж**
Czech language
užovka 1, 437 **вуж**
Slovak language
užovka 1, 437 **вуж**
Upper Lusatian language
wuž 1, 437 **вуж**
Lower Lusatian language
wuž 1, 437 **вуж**
Slovenian language
vóž 1, 437 **вуж**

Proto-Slavic language
*ǫžь 1, 437 **вуж**
Lithuanian language
angìs 1, 437 **вуж**
Prussian language
angis 1, 437 **вуж**
Latin language
anguis 1, 437 **вуж**
Old High German language
unc 1, 437 **вуж**
Middle Irish language
escung 1, 437 **вуж**
Indo-European language
*ang^{u}(h)i- 1, 437 **вуж**
…

The * sign is used for reconstructed words; we include it in the alphabet.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

2.4 Representation of the dictionary text in the structure of the lexicographic database

The current digital version of the dictionary uses a relational database of data. The entire corpus of dictionary entries is organized in 6 tabs. The tables **uketym_etym_classes**, **uketym_biblography**, **uketym_links** contain the texts of etymological classes from which the texts of dictionary entries are formed. In the **uketym_etymons** table, the parameters of all the explicated etymon structures are shown. The **uketym_languages_all** table contains information about the dictionary languages. The **ukretym_heads table** organizes the text of the dictionary entry. In the **uketym_etym_classes table**, the texts of etymological classes of dictionary entries (*HEAD*, *DER*, *SLAV*, *LANG*) are saved. All inter-article dictionary links are organized in **uketym_links**, all bibliographic links are in the **uketym_bibliography table**. The **uketym_language_all** table contains information about languages used in the dictionary. Based on this table, all user language registries are formed. The selection of an index array for a given language registry is performed from the **uketym_etymons table**.

The parsing of the dictionary text and the conversion of the text to the database was performed by a single program without the formation of intermediate files (in the future we abandoned this approach).

To perform parsing, the texts of dictionary entries of all volumes were converted into HTML format. The volumes of the dictionary were prepared for printing by various publishing technologies. The first three volumes are in monotype, pre-computer technology. Therefore, the printed texts were scanned and recognized using the ABBYY FineReader program, and then the texts were proofread. The last three volumes were prepared by means of a computer publishing system (MS Word was used). The sign system of the entire Dictionary text has been unified according to UNICODE 3.0 encoding. This made it possible to carry out an inventory of the alphabet symbols to represent the etymons of each language. To link between the printed and digital versions of the dictionary, each dictionary entry was marked (manually) as follows: volume number, page number at the beginning of the text of the dictionary entry, page number at the end of the text of the dictionary entry.

2.5 User interfaces

Basic VLL (virtual lexicographic laboratory) functions:

1) traditional entry by the head word and dictionary entry text visualization according to its structure (see fig. 1);
2) editing of any structural element of a dictionary entry;
3) building a dictionary entry of a given structure;
4) automatic indexing for each EDUL language (or a specified set of languages) (see fig. 2, fig. 3).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

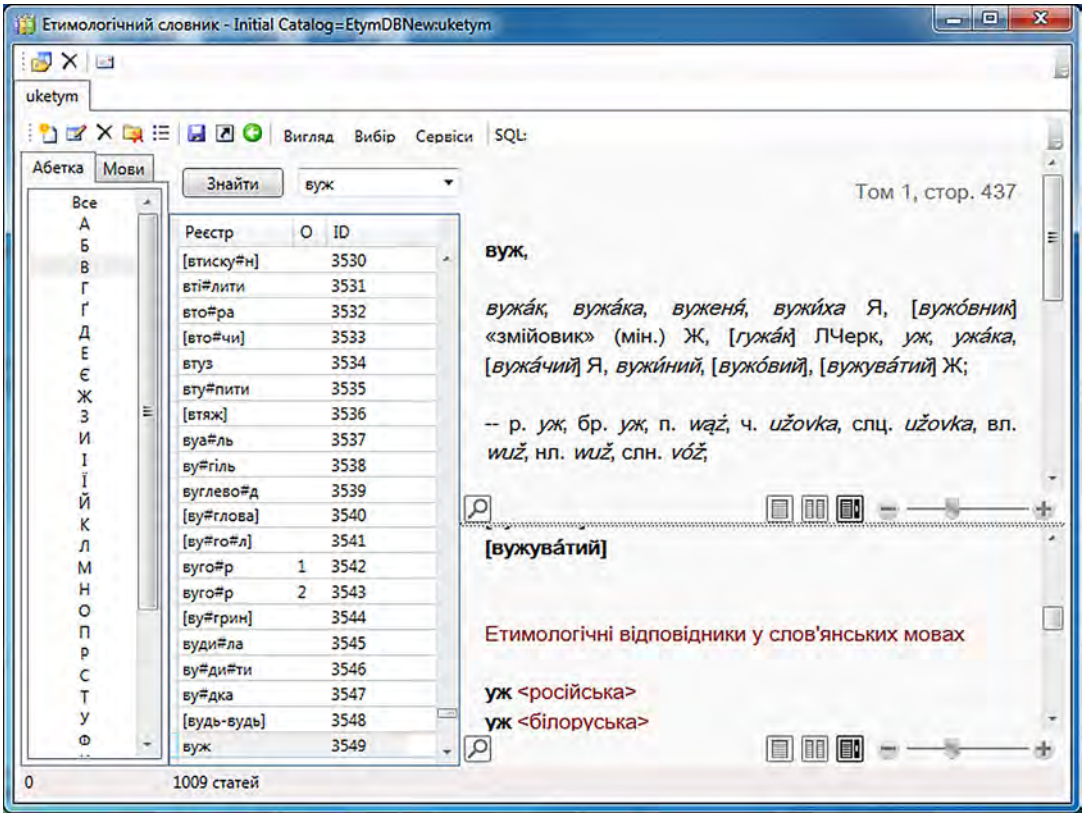


Fig. 1: Main window (article "**уж**")

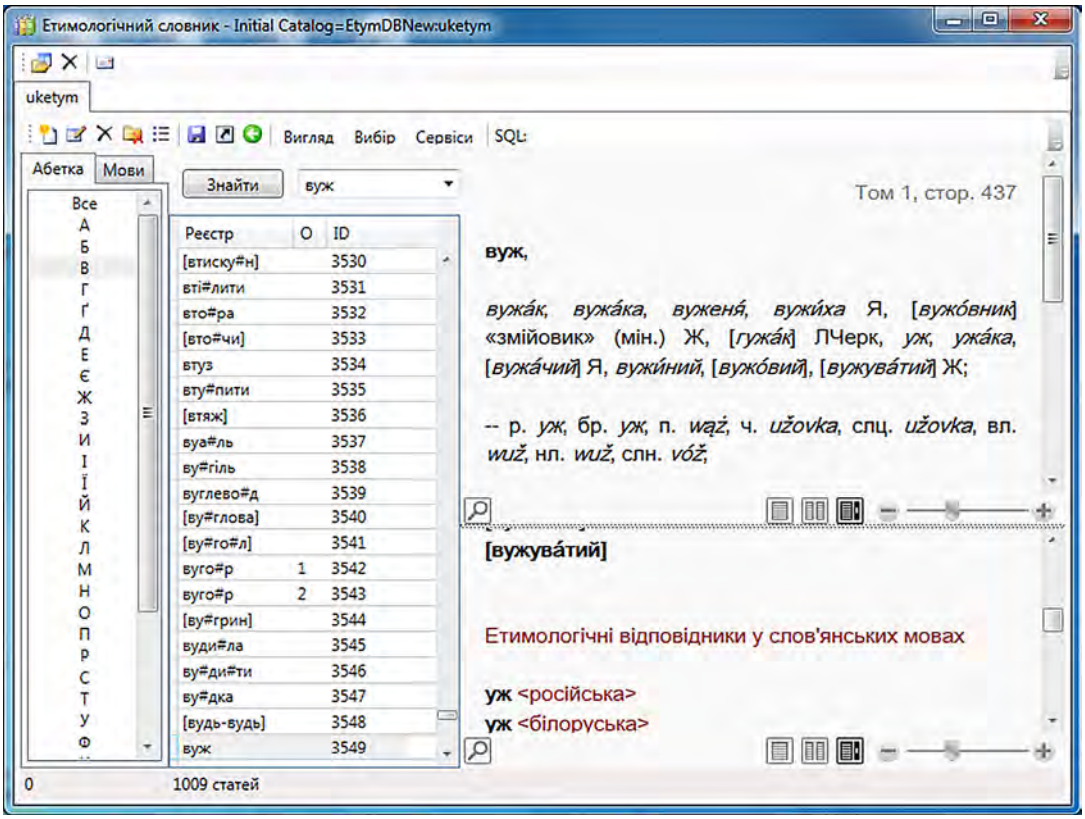


Fig. 2: Language index for a given language list (head words)

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

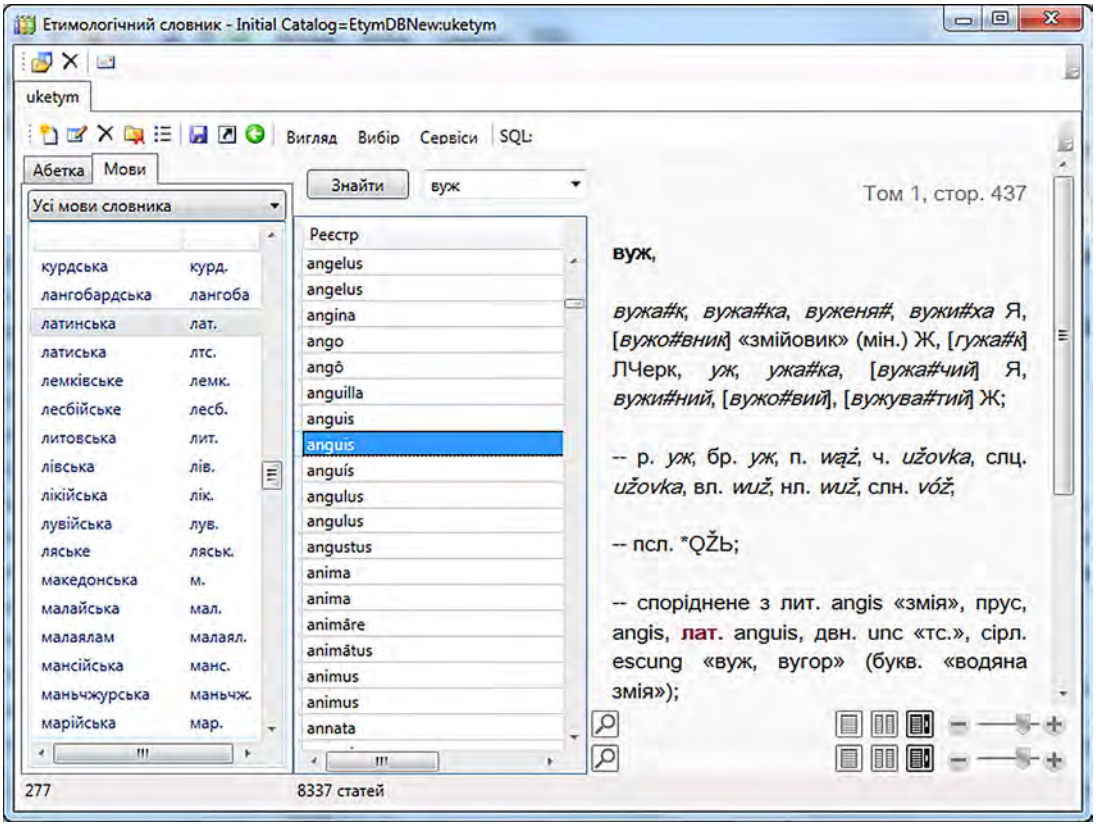


Fig. 3: Language index for a given language list (etymons)

3. Conclusion

In modern conditions, the sociologization of any lexicographic product is largely determined by its adaptation to functioning in the digital environment. This is true not only in relation to the mass consumer, but also in relation to professional communities. Such fundamental lexicographic projects as Dictionaries of national languages are switching (and some have already switched) to non-stop functioning mode: dictionary entries are updated and the registry is updated on a single lexicographic database with on-line user access.

Like most large dictionaries, etymological dictionaries also have their own representation in the digital environment. However, as a rule, they are presented in the format of machine-readable texts, which significantly limit their scalability and potential for research work.

EDUL has been created for half a century. The printed version of the index is understood more as the completion of a printed project, a tribute to tradition. Given more than modest print runs of the last three volumes (and the first three have already become a bibliographic rarity), the integrity of the dictionary and its sociologization can only be ensured by its digital version. To a large extent, the interest in the dictionary (in our understanding) will be provided by the development of the user interface, focused primarily on research tasks.

The new dictionary of the Ukrainian language (like its predecessor) does not have an etymological reference in its structure (SUM 2010–2021). The new edition of the dictionary is formed in the Digital Writing System mode and each new volume is delivered to users in the format of a printed book and a website on the Internet. Therefore, the integration of these two dictionaries is a useful technological task.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

References

ESUM (1982–2012): Etymologichnyĭ slovnyk ukraïskoĭ movy. V 7 t. (Vol. 1–6). Kyïv.

Shyrokov, V./Ostapova, I./Yakymenko, K. (2014): Digital lexicographical systems and traditional paper dictionaries (from traditional paper dictionaries to digital lexicographical systems). In: Cognitive Studies/Études cognitives 15, pp. 193–210.

Shyrokov, V./Ostapova, I. (2015): Indexing the etymological lexicographic systems. In: Cognitive Studies/Études cognitives 14, pp. 1–11.

Shyrokov, V (ed.) (2018): Computer linguistic studies: proceedings of the Ukrainian Lingua Information Fund NAS of Ukraine. Vol. V: Virtualization of linguistic technologies. https://movoznavstvo.org.ua/files/Ling_inf_studio_TOM_5_umif_B5.pdf (last access: 24-03-2022).

SUM (2010–2022): Slovnyk ukraïskoĭ movy v 20 t. Vol. 1–12. Ukrainian Lingua Information Fund NAS of Ukraine. Accesses at: https://services.ulif.org.ua/expl/, sum20ua.com (last access: 24-03-2022).

Trap-Jensen, L. (2018): Lexicography between NLP and linguistics: aspect of theory and practice. In: Čibej, J./ Gorjanc, V./ Kosem, I./Krek, S. (eds.): Lexicography in lobal Contexts. Proceedings of the 18th EURALEX International Congress 2018, 17–21 July 2018, Ljubljana. Ljubljana, pp. 25–38.

Contact information

Iryna Ostapova
Ukrainian Lingua-Information Fund of National Academy of Sciences of Ukraine
irinaostapova@gmail.com

Volodymyr Shyrokov
Ukrainian Lingua-Information Fund of National Academy of Sciences of Ukraine
Vshirokov48@gmail.com

Yevhen Kupriianov
National Technical University “Kharkiv Polytechnic Institute”
eugeniokuprianov@gmail.com

Mykyta Yablochkov
Ukrainian Lingua-Information Fund of National Academy of Sciences of Ukraine
gezartos@gmail.com

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Anna Pavlova

MEHRSPRACHIGE DATENBANK DER PHRASEM-KONSTRUKTIONEN

Abstract The paper describes an online German-Russian database for phraseological constructions (PhC), or syntactic idioms. It is a linguistic phenomenon representing a stable multi-word form that usually contains some auxiliary words ("anchors") and partially opens up empty spaces ("slots") which are filled directly in spoken language by various lexemes or combinations of lexemes ("fillers", or "slot fillers"). Linguists from several German institutions are currently working on the database. The PhCs selected for the database have to meet special criteria. The database is a manual that combines scientific descriptions, a thesaurus and a bilingual dictionary. The database is designed as an active aid for text production in the respective foreign language; it is also a manual for language researchers and for translators. Apart from that, it can serve as a basis for extensions for other language pairs. The aim of the project is to record and to describe 300 PhC before the database is published. Our objective is to enable foreign language learners to use the syntactic idioms correctly in the texts they produce rather than create a big-sized database. The paper describes some issues related to the creation of the database, namely objectives and target groups, material and methods, microstructure of the database article and some others.

Keywords Phraseme constructions; syntactic phrasemes; syntactic idioms; online database; bilingual dictionary

1. Einleitung

Der Fachbegriff „Phrasem-Konstruktion" stammt von Dmitrij Dobrovol'skij.[1] Die Phrasemkonstruktionen können

> [...] als Konstruktionen definiert werden, die als Ganzes eine lexikalische Bedeutung haben, wobei bestimmte Positionen in ihrer syntaktischen Struktur lexikalisch besetzt sind, während andere Slots darstellen, die gefüllt werden müssen, indem ihre Besetzung lexikalisch frei ist und nur bestimmten semantischen Restriktionen unterliegt. (Dobrovol'skij 2011, S. 114)

Es handelt sich somit um ein sprachliches Phänomen, das eine stabile Mehrwortform darstellt, die in der Regel teilweise gefüllt ist („Anker") und teilweise Leerstellen („Slots") eröffnet, die direkt in der gesprochenen Sprache durch verschiedene Lexeme oder Kombinationen von Lexemen („Filler", oder „Slotfüller") gefüllt werden. Durch das Füllen der Slots entsteht bei der Textproduktion ein Konstrukt, das auf einem der Ebene des Sprachsystems angehörenden Modell basiert. Phraseologische Konstruktionen (PhK) gehören somit zu den sprachlichen Phänomenen, die Sprache als System (langue) und Sprache als Tätigkeit (parole) miteinander verknüpfen. Solche Konstruktionen sind idiomatisch, da ihre Bedeutung nicht aus der Summe der grammatikalischen und lexikalischen Bedeutungen ihrer Bestand-

[1] Es gibt auch andere Bezeichnungen für dieses Phänomen. Von Wolfgang Fleischer stammt der Fachbegriff „Phraseoschablone" (Fleischer 1982, S. 136). In der sowjetischen Linguistik hat Natalia Švedova als Erste diese Konstruktionen beschrieben; sie hat ihnen den Namen „phraseologisierte Gebilde" („frazeologizirovannye postroenija") gegeben (1960, S. 279). Dmitrij Šmelëv führte in die sowjetische Sprachwissenschaft den Fachbegriff „frazeoschema" ein (1977, S. 327–330). Alla Veličko nennt dasselbe Phänomen „syntaktische Phraseologismen" (1996). Igor Mel'čuk beschreibt sie als „syntactic phrasemes" (Mel'čuk 1987) und „syntactic idioms" (Mel'čuk 2021). Leonid Iomdin spricht in diesem Zusammenhang von der „Mikrosyntax" (2006).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

teile ableitbar ist. Um die Ausdrücke *Da habt ihr eure Hochzeit!* oder *Da hast du deinen Traum!* angemessen zu verstehen, reicht es also nicht aus, die grammatikalischen Formen und lexikalischen Bedeutungen der im Konstrukt beteiligten Wörter zu kennen: Man muss die Konstruktion als Ganzes im Voraus als Modell (PhK) kennen, denn sie verfügt über eine Bedeutung, die sich nicht aus ihren Elementen und deren Summe ableiten lässt. Auch wenn die aufgefüllten Slots in einem anderen Kontext anders aussehen würden, hat das Modell für sich eine idiomatische Bedeutung, die sich definieren und beschreiben lässt. Im Fall *Da habt ihr*/*Da hast du* [*dein*/*euer*] *X!* ist es das Vorweisen eines Objektes X, meist begleitet durch Skepsis, Hohn oder ähnliche illokutive Bedeutungen.

> Konstruktionen dieser Art haben eine festgeprägte Modellbedeutung, die bei Ausfüllung des Modells mit entsprechendem lexikalischem Material eine Wortverbindung erzeugt, deren allgemeine Bedeutung durch die Bedeutung des Modells bereits vorbestimmt ist. (Fleischer 1982, S. 136)

Trotz Erwähnungen und (relativ knappen) Beschreibungen einzelner PhK in den 70er und 80er Jahren des vorigen Jahrhunderts vor allem als einer Art Randerscheinung im Bereich Phraseologie, rückten die PhK als eigenständiges Thema in den Fokus der Sprachwissenschaft erst mit dem wachsenden Interesse an der Konstruktionsgrammatik (CxG, siehe Goldberg 1995). Es erwies sich, dass PhK unter Konstruktionen einen wichtigen Platz einnehmen (vgl. Schafroth 2014; Finkbeiner 2017; Janda/Kopotev/Nesset 2020; Pavlova 2020).

Nach Fillmore/Kay/O'Connor (1988) liegen alle sprachlichen Konstruktionen in einem Kontinuum zwischen zwei Polen: dem der lexikalischen Offenheit (nichtidiomatische Konstruktionen) und dem der lexikalisch fertig und fest gefüllten Konstruktionen (formelartigen Sätzen). Laut Goldberg (2006, S. 5):

> Any linguistic pattern is recognized as long as some aspect of its form or function is not strictly predictable from component parts or from other constructions recognized to exist. In addition, patterns are stored as constructions even if they are fully predictable as long as they occur with sufficient frequency.

Die rasante Entwicklung der CxG (s. dazu auch Fischer/Stefanowitsch 2008; Ziem 2018) bewegt viele Linguisten weltweit dazu, nach den Wegen der Abbildung und der möglichst ausführlichen Beschreibung verschiedener grammatischer Konstruktionen in der Lexikografie zu suchen. Eines der ersten Projekte dieser Art stellt das Konstruktikon dar (Janda et al. 2018), das über 2500 Einträge beinhaltet und in der Abschlussphase begriffen ist. Die PhK bilden nur einen kleinen Teil der Makrostruktur des Konstruktikons.

Unser Ansatz ist ein anderer. Es wird zurzeit an zwei zweisprachigen Online-Datenbanken für phraseologische Konstruktionen gearbeitet, einer russisch-deutschen und einer deutsch-russischen. Die russisch-deutsche Datenbank wird von einem Team von Sprachwissenschaftlern aus Moskau, Jekaterinburg, Germersheim (Deutschland), Jerusalem und Stockholm erstellt. An der deutsch-russischen Datenbank arbeiten Sprachwissenschaftler aus mehreren deutschen Instituten und Universitäten.[2] Beide Gruppen entwickeln die Datenbanken aufgrund der gleichen Vorlage, der gleichen Online-Plattform und den gleichen Methoden. Einige Einträge in der russisch-deutschen Datenbank werden mit semantisch ähnlichen Einträgen in der deutsch-russischen Datenbank verknüpft und umgekehrt. Darüber hinaus sind in vielen Datenbank-Artikeln in der russisch-deutschen Datenbank Verweise

[2] Im Moment wird darüber diskutiert, für die deutsche Datenbank mehr Übersetzungssprachen als nur Russisch aufzunehmen. Im Gespräch sind Polnisch, Tschechisch, Serbisch, Slowenisch, Kroatisch, Ukrainisch. Die deutsche PhK Datenbank wird somit zu einer mehrsprachigen ausgebaut.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

auf das Konstruktikon (Janda et al. 2018) enthalten, die wirksam werden, sobald diese Ressource veröffentlicht wird.

Die für die Datenbanken ausgewählten PhK müssen folgenden Kriterien entsprechen: 1) sie verfügen über ihre eigene Bedeutung, noch bevor sie im Text aktuell aufgefüllt werden; diese Bedeutung lässt sich beschreiben; 2) PhK sind idiomatisch, d.h. ihre Bedeutung ist nicht transparent und kann nicht direkt aus deren lexikalischen Bestandteilen und ihrer Morphologie abgeleitet werden; 3) die Anker sind meist Synsemantika, aber es gibt eine Reihe von Ausnahmen; diese werden mit berücksichtigt; 4) die PhK müssen als Modelle produktiv sein, d. h. es muss eine gewisse Freiheit bei der Slot-Auffüllung bestehen, auch wenn diese von PhK zu PhK unterschiedlich ist und gewissen Restriktionen unterliegt; 5) PhK unterscheiden sich von konstitutionellen Sätzen und Phrasen durch ihre Prosodie, die für PhK festgelegt ist; 6) PhK erfüllen im Text verschiedene syntaktische Funktionen (Subjekt, Prädikat, selbständiger Satz, Nebensatz etc.); 7) die meisten PhK gehören zur Umgangssprache, aber es gibt auch hier Ausnahmen; 8) PhK sind stark im Text eingebunden; im Text bekommen sie zusätzliche Bedeutungen, die aber nicht der Hauptbedeutung des Modells widersprechen, sondern sie bereichern; 9) PhK sind idiosynkratisch: Sie sind nicht grammatikalisiert und können nicht als reguläre Ausdrucksformen für grammatische Kategorien dienen.

Die beiden Datenbanken sind als Nachschlagewerke konzipiert, die eine Kombination aus Lexikon enzyklopädischen Typs und zweisprachigem Wörterbuch darstellen.

Im Weiteren werden einzelne Aspekte des Projektes am Beispiel der deutsch-russischen PhK-Datenbank (abgekürzt PhK-DB, oder einfach DB) beleuchtet.

2. Ziele und Zielgruppen

Die DB ist für drei Hauptzielgruppen konzipiert: 1) Für die Fremdsprachenlernenden und -lehrenden, d.h. für didaktische Zwecke. Dabei geht es darum, nicht nur Informationen über die PhK zu vermitteln, sondern durch Beschreibungen und Beispiele die Lerner zu befähigen, die PhK in der Fremdsprache aktiv und korrekt zu gebrauchen: Es liegt dem Projektteam viel daran, eine Art aktives Wörterbuch zu erstellen. Die didaktischen Zwecke liegen auch dem Ansatz zugrunde, die PhK einerseits möglichst ausführlich zu beschreiben, andererseits nicht zu viele linguistische Fachbegriffe und nicht zu lange Sätze in der Mikrostruktur zu benutzen. 2) Für Sprachwissenschaftler, die sich mit den Themen Phraseologie und Konstruktionsgrammatik befassen. Auch für Sprechwissenschaftler wäre unser Projekt von Interesse, denn einer der darin abgebildeten Aspekte ist Prosodie der PhK. 3) Für Übersetzer. Deshalb nimmt in der Mikrostruktur der Pool von Übersetzungsbeispielen und Kommentaren zu jedem Beispiel eine zentrale Stelle ein. Eine Herausforderung für Projekt-Teilnehmer ist es, eine passende Definition der Äquivalenz für die Beschreibung der Konstrukt-Übersetzungen zu finden.

3. Makrostruktur und Mikrostruktur

Die Datenbank entsteht auf der Online-Plattform github. Die Datenbank als Ganzes beinhaltet ein Vorwort, eine Liste der Abkürzungen, eine Liste der verwendeten Terminologie mit Erklärungen und die eigentliche DB.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Die Makrostruktur stellt eine Liste der Lemmata in Form von alphabetisch geordneten PhK dar, z. B. *Da hast du N; Dass du ja nicht V; Du bist mir ein schöner N* etc. Die äußere Selektion erfolgt nach den Kriterien für PhK, wie sie in Dobrovol'skij (2011, 2016) definiert und in der Einleitung zu diesem Aufsatz beschrieben sind. Die innere Selektion (Wahl, Form und Ausführlichkeitsgrad der in der Mikrostruktur dargebotenen lexikographischen Informationen) richtet sich nach den Bedürfnissen der oben genannten Zielgruppen. Somit ergeben sich folgende Aspekte (Tags) für die Mikrostruktur:

<function>	Beschreibung des semantischen Feldes auf einer abstrakten Ebene (z. B. Topikalisierung, Bewertung, Identität etc.)
<lemma>	Das PhK-Modell als Formel, z. B.: *N'$_{nom}$ ist/sind nicht gleich N'$_{nom}$* (*Kind ist nicht gleich Kind*).
<variants>	Eine bis mehrere Varianten der Lemma-PhK. Unter Varianten versteht man die sich der Lemma-PhK morphologisch und semantisch ähnelnden Modelle mit kleineren formellen Abweichungen. So wäre z. B. eine Variante zu einer bestimmten PhK diejenige, in der eine andere Wortfolge gegenüber der Lemma-PhK möglich wäre.
<meaning>	Beschreibung der PhK-Semantik. Wird für eine PhK eine Mehrdeutigkeit festgestellt, wird in diesem Tag vermerkt, dass dies Bedeutung 1, Bedeutung 2 etc. ist.
<examples_for_lemma>	Einige Beispiele aus verschiedenen Textkorpora, ohne Übersetzung.
<morpology>	Beschreibung der Slotfüllungen aus Sicht ihrer morphologischen Eigenschaften.
<syntax>	Syntaktische Rolle (oder Rollen) des Konstruktes im Satz und im Text (z. B. Prädikativ, eine selbstständige Proposition, Temporalsatz etc.)
<usage>	In welchen Textsorten und Texttypen kann das Konstrukt verwendet werden? Prüfen, ob es als Anfangssatz oder nur als Antwortreplik möglich wäre etc.
<style>	Stilistisches Register wird ausführlich beschrieben.
<prosody>	Eine detaillierte Schilderung prosodischer Merkmale; danach folgt eine Audio-Aufnahme eines Beispiels. Bei komplexen prosodischen Konturen von PhK werden auch F0-Grafiken (Grundtonverlauf) beigefügt.
<examples_translation>	Dieser Teil ist in Form einer Tabelle abgebildet, deren Spalten „Originalbeispiel", „Übersetzung" und „Kommentar zur Übersetzung" heißen. Übersetzungen werden den parallelen Textkorpora (das Parallelkorpus des Nacional'nyj korpus russkogo jazyka; Weblitera) oder der Literatur in Form von konventionellen Büchern (Original – Übersetzung) entnommen. Bei fehlenden publizierten Übersetzungen erstellen Projektteilnehmer eigene Übersetzungen (das sind aber Ausnahmefälle).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

<general comment>	Abschließende und verallgemeinernde Kommentare zu den Übersetzungsoptionen (Äquivalenz vorhanden oder nicht vorhanden, welche Übersetzungsoptionen werden am häufigsten beobachtet etc.).
<formulaic_expressions_build_on_PhK>	Manchmal gibt es in der Sprache fertige Formelsätze, die auf der Lemma-PhK oder deren Varianten gründen und sich mit der Zeit verfestigt haben.
<idioms_build_on_PhK>	Es kann auch traditionelle (lexikalische) Phraseme geben, denen das PhK-Modell zugrunde liegt.
<addional info>	Liegen zur PhK zusätzliche Informationen vor (z. B. historische Entwicklung, Häufigkeitsindikatoren), werden sie in diesem Tag erfasst.
<full synonyms>	Zu manchen PhK gibt es volle Synonyme (kommt selten vor).
<quasi synonyms>	Die meisten Synonyme für PhK sind Quasi-Synonyme, d.h. syntaktische Modelle mit einer ähnlichen oder sogar gleichen Bedeutung, die aber nicht in jedem Kontext durch die Lemma-PhK ausgetauscht werden können.
<homonym>	Für manche PhK können Homonyme entdeckt werden, d. h. gleich aussehende Modelle bzw. konstitutionelle Phrasen mit einer anderen Bedeutung.
<paronym>	Paronyme, genauso wie Homonyme, sind nicht für jede PhK auffindbar.

Diesem ausführlichen Profil und mehreren Beispielen und Kommentaren folgen Links auf die PhK der deutschen Datenbank, die ähnlich aufgebaut sind bzw. eine ähnliche Bedeutung haben. Auch auf die PhK aus der russischen Datenbank wird verwiesen, die für die Lemma-PhK als Äquivalente infrage kämen. Bei mehrdeutigen PhK wird die Mikrostruktur um die nötigen Tags erweitert.

So etwa sieht ein konkretes Beispiel aus:[3]

<function> Beständigkeit, Beurteilung

<lemma> **N' nom ist/sind [immer] N' nom**

<meaning> Aus Sicht formaler Logik ist das Konstrukt eine tautologische Äußerung. Doch linguistisch hat sie einen nichttautologischen Sinn: Das Objekt, das im ersten Slotfüller genannt wird, wird durch die Konstruktion in eine Klasse der gleichartigen und gleichnamigen Objekte mit einbezogen und gleichzeitig wird darüber ausgesagt, dass all diese Objekte durch eine bestimmte Eigenschaft vereint sind. Außerdem wird mitgeteilt, dass diese Eigenschaft über einen hohen Grad Beständigkeit verfügt, sie bleibt über eine längere Zeit unverändert. Die Eigenschaft selbst wird dabei nicht benannt; der Adressat muss selbst darauf kommen. Die PhK enthält für den Adressaten eine Art Rätsel. Auch Eigennamen sind als Slotfüller möglich. In diesem Fall handelt es sich nicht um die Einbeziehung in eine Klasse von ähnlichen Gegenständen, sondern lediglich darum, dass sich eine konkrete Eigenschaft (oder evtl. auch mehrere) des durch den Eigennamen bezeichneten Denotats über eine längere Zeit konstant bleibt; man dürfte hierfür keine Veränderungen erwarten. Das Konstrukt wird oft als Abmahnung verwendet: Jemand will etwas rechtfertigen, kleinreden oder vergessen, was aus Sicht des Sprechers nicht rechtfertigt, kleingeredet oder vergessen werden darf, da das im Slotfüller genannte Objekt über bestimmte Eigenschaften verfügt, die es nicht erlauben. Das Konstrukt kann

[3] Änderungen sind vorbehalten.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

aber auch zum Trost verwendet oder mit anderen Illokutionen versehen werden. Eine semantisch und strukturell ähnliche PhK ist **N' nom bleibt/bleiben N' nom** .

<examples_for_lemma> ***Waffe ist Waffe. Tod ist Tod***. (Welt.de); ***Krieg ist Krieg***. *Es gibt nichts Schlimmeres.* (infosperber.ch); *Elite weiß nicht, was das Proletariat denkt.* ***Elite ist Elite***. *Man kann sich nicht selbst zur Elite zählen, man kann nicht Elite sein wollen.* ***Elite bleibt Elite***. *Elite respektiert nur Elite.* (Demian Varth. Brainstorming); *Ich bin da eher auf der Seite der Hühner.* ***Winter ist Winter, Dunkelheit ist Dunkelheit.*** *Ich würde im Winter auch gern länger schlafen.* (Wiener Zeitung); *Irgendwann musste ich einsehen:* ***Die Schweiz ist die Schweiz und Brasilien ist Brasilien***. (tagblatt.ch/kultur)

<morpology> Der Slotfüller ist ein Substantiv im Nominativ, das wiederholt wird. Normalerweise wird das Verb *sein* im Präsens Indikativ verwendet, in Form von *ist* oder *sind*. Auch Eigennamen sind als Slotfüller möglich.

<semantic restrictions> Es sind keine semantischen Einschränkungen bekannt. Der Slotfüller muss aber so gewählt werden, dass der Adressat verstehen würde, welche Eigenschaften des entsprechenden Objektes gemeint sind, besonders wenn sich im Kontext keine Hilfen finden lassen.

<syntax name> Selbstständiger Satz. Der erste Slotfüller ist Subjekt des Satzes, der zweite ist Prädikativ.

<usage> Im Narrativ oder im Dialog, im mündlichen oder schriftlichen Diskurs, in fast allen Textsorten außer denen, die solche philosophisch anmutenden Sentenzen nicht zulassen (z. B. technische Doku oder juristische Dokumente).

<style> Neutral bis gehoben, rhetorisch, philosophisch geprägt.

<prosody> Ein Doppel-Akzent-Konstrukt: ein starker Akzent markiert beide Slotfüller[4].

<examples_translation>

Beispiel	Übersetzung	Kommentar
„Mein Freund", rief ich aus, „**der Mensch ist Mensch**, und das bißchen Verstand, das einer haben mag, kommt wenig oder nicht in Anschlag, wenn Leidenschaft wütet und die Grenzen der Menschheit einen drängen." (Goethe. Die Leiden des jungen Werther)	«Друг мой! — вскричал я. — **Человек всегда останется человеком**, и та крупица разума, которой он, быть может, владеет, почти или вовсе не имеет значения, когда свирепствует страсть и ему становится тесно в рамках человеческой природы». (Üb. Наталия Касаткина, 1954)	Im russischen Äquivalent wird entweder das Kopula-Verb *быть* oder das Verb *оставаться/остаться* verwendet.

[4] Prosodische Beschreibung wird in der Endversion etwas anders und ausführlicher ausfallen und durch eine Audio-Aufnahme ergänzt.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Beispiel	Übersetzung	Kommentar
„Greift zu“, sagte Gullik, der es freundlich meinte, „es ist das Beste. Wort ist Wort.“ (Theodor Mügge. Afraja)	«Соглашайтесь, — настаивал Гуллик, настроенный миролюбиво, — никто вас обманывать не собирается». (Eigene Übersetzung, 2022)	Das Konstrukt *Wort ist Wort* kann nicht durch ein russisches Äquivalent übersetzt werden. Denn das russische Konstrukt *Слово есть слово* wird verwendet, wenn der Adressat dabei ist, sein Wort zu brechen, oder wenn der Sprecher selbst daran zweifelt, dass das von ihm gegebene Wort eingehalten werden müsste. Und im deutschen Text ist es anders gemeint: Das ist eine Beruhigung für diejenigen, denen gerade ein Vorschlag gemacht wurde. Da würde das russische äquivalente Konstrukt nicht passen.
Der Kahlwicht ist einverstanden. **Gesetz ist Gesetz**. Er wird sich nicht dagegen sträuben. (Erwin Strittmatter. Tinko)	Лысый чёрт согласился. **Закон есть закон**. Не будет же Лысый чёрт возражать против законов! (Üb. Всеволод Розанов, 1956)	Das Äquivalent im Russischen ist vom Typ „eins zu eins“.
„Sehr groß ist der Unterschied nicht, zwischen dem, was ich einmal war, und dem, was ihr nun seid. Denn **Kind ist Kind**, und **Schule ist immer Schule**.“ (Galsan Tschinag. Die graue Erde)	«Разница невелика, и я был когда-то таким, как вы теперь. Ведь **дети есть дети, а школа есть школа**». (Eigene Übersetzung, 2022)	Das Äquivalent im Russischen ist vom Typ „eins zu eins“.
„Komm, Sepp! Es hilft nichts. **Dienst ist Dienst!** Wir müssen dich abschieben!“ (Erich Maria Remarque. Liebe Deinen Nächsten)	«Пошли, Зепп! – сказал он Штайнеру. – Ничего не попишешь! **Служба есть служба**! Мы обязаны переправить тебя!» (Üb. Исаак Шрайбер, 1990)	Das Äquivalent im Russischen ist vom Typ „eins zu eins“.
Дети есть дети, сидим на уроке, в «морской бой» играем. (С. А. Алексиевич. Последние свидетели)	**Kinder sind Kinder** – wir saßen im Unterricht und spielten „Schiffe versenken“. (Üb. Ganna-Maria Braungardt, 2016)	Im Russischen ist eine identische Konstruktion vorhanden.

<general translation comment> Im Russischen existiert ein volles Äquivalent vom Typ „eins zu eins“, sowohl strukturell als auch stilistisch: **N’ nom есть N’ nom**.

<synonyms> PhK **N’ nom bleibt/bleiben N’ nom** : *„Die Sonne dreht sich um die Erde“ ist keine Meinung, sondern schlicht und einfach falsch:* ***Fakten bleiben Fakten***. (Susanne Schnabl. Wir müssen reden: Warum wir Streitkultur brauchen); PhK **N’ nom ist und leibt/sind und bleiben N’ nom:**

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Ein Verbrechen ist und bleibt ein Verbrechen (religion.orf.at); ***Kinder sind und bleiben Kinder*** (amazon.de/Kinder-gestern-heute-Nina-Mielke). Ein volles Synonym sind diese PhK jedoch nicht. Vgl.: *Clevere und smarte Mitarbeiter existieren in allen Unternehmen, jedoch wird nur selten das volle organisatorische Potenzial ausgeschöpft.* ***Ideen bleiben Ideen, Visionen bleiben Visionen.*** (mindsandelephant.com). Hier kann die Konstruktion nicht durch die Lemma-PhK ersetzt werden.

<full synonyms>

<addional info>

<idioms_build_on_PhK> ***Dienst ist Dienst, und Schnaps ist Schnaps.***

<formulaic_expressions_build_on_PhK">

<homonym>

<paronym> Kompositionelle Sätze, in denen das Subjekt und das Prädikativ durch dasselbe Wort ausgedrückt sind. Das rekurrente Wort wird durch einen Attributsatz ergänzt: *Modifizierte Stärke ist Stärke, die technologisch so verändert ist, dass sie als Bindemittel eingesetzt werden kann.* (keuneckefeinkost.de)

Somit wird ein möglichst volles Bild (ein ganzheitliches Profil) der PhK angestrebt, unter Berücksichtigung der Deutsch-als-Fremdsprache-Lerner und der angehenden Übersetzer.

4. Datenerhebung: Material und Methoden

Als Hauptquelle für die Erstellung der deutschen DB dient eine Liste der deutschen PhK (ca. 150), die von einer Projekt-Teilnehmerin seit geraumer Zeit (ca. 5 Jahre Vorlaufzeit) in der Excel-Form anhand von Gehörtem und Gelesenem erstellt wurde. Dazu kommen die PhK, die während der aktuellen Arbeit am Projekt weiterhin durch Beobachtungen der gesprochenen und geschriebenen Texte gefunden werden. Als Material dienen mündliche und schriftliche Texte[5], die die Projektteilnehmer wahrnehmen und analysieren und in denen sie nach den PhK suchen. Als dritte Quelle dienen Übersetzungen russischer PhK, denn gerade dank Übersetzungen werden einige deutsche PhK als Äquivalente zu den russischen PhK entdeckt (und umgekehrt). Nach Beispielen wird in den Textkorpora gesucht, und zwar im Parallel-Unterkorpus des Nacional'nyj korpus russkogo jazyka, DWDS, DeReKo und Sketch Engine. Auch Projekt Gutenberg wird intensiv genutzt. Als weitere Quellen für die Datenerhebung dienen konventionelle Bücher und Zeitschriften, Werbung, Blogs, Internet-Foren. Die gefundenen Beispiele dienen ihrerseits als Grundlage für die Kenntnisse, die wir über morphologische, semantische, stilistische, prosodische Beschaffenheiten der PhK gewinnen, so dass die einzelnen DB-Artikel immer wieder durch neue Erkenntnisse durch Beispiel-Analysen ergänzt werden.

5. Umfang

Es ist angestrebt, insgesamt 300 PhK zu erfassen. Danach wird die DB veröffentlicht. Die DB erhebt also keinen Anspruch auf einen möglichst großen Umfang; uns liegt viel mehr daran, die PhK ausführlich zu beschreiben und somit die Fremdsprachenlerner zu befähigen, diese in den von ihnen produzierten Texten korrekt einzusetzen. Auch für die angehenden Über-

[5] Bei mündlichen Texten handelt es sich um Alltagsdialoge, Filme, Serien, Radiosendungen, Youtube-Streams.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

setzer ist es wichtig, möglichst ausführliche Informationen über die Optionen zu erhalten, die ihnen bei der Arbeit an Texten zur Verfügung stehen, wo sie PhK begegnen.[6]

6. Wissenschaftliches Potenzial

Für die Sprachwissenschaft eröffnen sich durch die hier schematisch vorgestellte DB mehrere neue Themenbereiche und Forschungsrichtungen: Geschichte einzelner PhK, ihre diachronische Entwicklung, Forschung von Illokutionen und Illokutionstypen, Validierung der Gebrauchshäufigkeit, Variabilität, semantische Eigenschaften von Fillern, Mehrdeutigkeit im Bereich PhK, prosodische Homonymie, Kreativität im Zusammenhang mit CxG und viele andere seien nur als Beispiele möglicher Forschungsrichtungen genannt. Im Bereich Morphologie bieten sich die Möglichkeiten zur detaillierten Forschung von Kombinierbarkeit. Lässt sich als Filler z. B. ein Substantiv feststellen, wäre zu untersuchen, ob nur Gattungsnamen oder auch Eigennamen zugelassen sind, ob ein Einzelsubstantiv oder eine Substantivgruppe als Filler fungieren können, ob neben konventionellen Substantiven auch substantivierte Adjektive oder substantivierte Infinitive infrage kämen etc. Auch Translationswissenschaft kann dank der DB zu neuen Erkenntnissen gelangen, wie z. B. Verfeinerung des Konzeptes Äquivalenz, Untersuchung der möglichen Ursachen für falsche Übersetzungen (dafür gibt es Belege), Erweiterung der Erkenntnisse über gelungene Übersetzungsentscheidungen und Übersetzungsverfahren, Untersuchung der Rolle von PhK für die Textqualität und den Stil. Auch für die Translationsdidaktik und für die Fremdsprachendidaktik bietet die DB neue Anregungen. Indirekt könnte auch die Korpuslinguistik vom neuen Projekt profitieren, denn anhand der Hürden, die man als Ersteller dieser DB zurzeit überwinden muss, um auf Beispiele zu kommen, wäre vorstellbar, dass die Suchmaschinen für einzelne Textkorpora für diese Art Aufgaben angepasst werden könnten.

7. Einige Schwierigkeiten und Herausforderungen

Es seien hier nur einige von vielen Problemen und Herausforderungen exemplarisch genannt. Eine Herausforderung besteht darin, die PhK einzugrenzen, sie auf dem Hintergrund benachbarter und ähnlicher Erscheinungen auszumachen. Ausgehend von der Definition der PhK sollte man dieses Phänomen von grammatischen Konstruktionen unterscheiden, die stark grammatikalisiert sind und im Rahmen des Fremdsprachunterrichts im Kurs Grammatik gelernt werden. Eine andere benachbarte sprachliche Erscheinung ist ein metaphorisch verwendetes Verb mitsamt Valenzpositionen. Auch Wortgruppen, die als „traditionelle“ Phraseologismen gelten, könnten als Lemma-Kandidaten erachtet werden, wenn man ihre Valenzstellen als Slots betrachten würde. Es sollten in diesem Zusammenhang auch Sprachformeln bedacht werden, die nach dem gleichen syntaktischen Modell aufgebaut sind, das aber in der modernen Sprache nicht (mehr) produktiv ist.

Auch die Bestimmung von Polysemie oder Homonymie der PhK stellt eine Herausforderung dar.

[6] Zu beobachtende Übersetzungsfehler in den veröffentlichten übersetzten (vor allem literarischen) Texten sind Belege dafür, dass PhK sowohl im Fremdsprachenunterricht als auch im Übersetzungs- und Dolmetschunterricht extra angegangen werden sollten.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Ein weiteres Problem besteht darin, passende Beispiele zu finden. Es ist nicht immer einfach, besonders wenn die PhK mit einem Filler beginnt. Sind genug Beispiele gesammelt, steht schon die nächste Herausforderung vor den Autoren, und zwar die Suche nach Übersetzungen.

8. Ausblick und Perspektiven

Im Moment wird das Projekt zur Erstellung der deutschen DB durch weitere slavische Sprachen erweitert. Somit würde sich die zweisprachige Datenbank in eine mehrsprachige verwandeln. Eine weitere wertvolle Ergänzung für das Gesamtprofil der PhK wäre, die Frequenz für den Gebrauch der PhK als Ganzes einerseits und für die Filler andererseits bemessen zu können. Es ist schwierig, Statistiken über phraseologische Konstruktionen aus Korpora zu erhalten, da die Füllung der Slots stark variiert. Dennoch kann man versuchen, solche Statistiken mindestens für einige PhK zu erstellen. Eine Projektteilnehmerin ist gerade dabei, die Methodik für die Analyse der Frequenz zu entwickeln. Die DB kann als Vorlage für ähnlich konzipierte Nachschlagewerke für weitere Sprachpaare dienen.

Literatur

Boas, H. C. (2019): Zur methodologischen Grundlage der empirischen Konstruktikographie. In: Czicza, D./Dekalo, V./Diewald, G. (Hg.): Konstruktionsgrammatik VI. Varianz in der konstruktionalen Schematizität. Tübingen, S. 237–263.

Dobrovol'skij, Dmitrij (2011): Phraseologie und Konstruktionsgrammatik. In: Lasch, A./Ziem, A. (Hg.): Konstruktionsgrammatik III. Aktuelle Fragen und Lösungsansätze (= Stauffenburg Linguistik 58). Tübingen, S. 110–130.

Dobrovol'skij, Dmitrij (2016): Grammatika konstrukcij i frazeologija. In: Voprosy jazykoznanija 3, S. 7–21.

Fillmore, Ch./Kay, P./O'Connor, M. C. (1988): Regularity and idiomaticity in grammatical constructions: the case of let alone. In: Language 64, S. 501–538.

Finkbeiner, R. (2017): "Argumente *Hin*, Argumente *Her*". Regularity and idiomaticity in German *N Hin*, *N Her*. In: Journal of Germanic Linguistics 29 (3), S. 205–258. https://doi.org/10.1017/S1470542716000234 (Stand: 23.3.2022).

Fischer, K./Stefanowitsch, A. (2008): Konstruktionsgrammatik: Ein Überblick. In: Fischer, K./Stefanowitsch, A. (Hg.): Konstruktionsgrammatik I: Von der Anwendung zur Theorie. Tübingen, S. 3–17.

Goldberg, A. E. (1995): A construction grammar approach to argument structure. Chicago/London.

Goldberg, A. E. (2006): Constructions at work: the nature of generalization in language. Oxford.

Iomdin, L. (2006): Mnogoznačnye sintaksičeskie frazemy: meždu leksikoj i sintaksisom. In: Laufer, Natal'ja/Narin'ani, Aleksandr/Selegej, Vladimir (Hg.): Komp'juternaja lingvistika i intellektual'nye technologii. Moskau, S. 202–206.

Janda, L./Lyashevskaya, O./Nesset, T./Rakhilina, E./Tyers, F. M. (2018): A constructicon for Russian: filling in the gaps. In: Lyngfelt, B. et al. (Hg.): Constructicography: constructicon development across languages. Amsterdam/Philadelphia, S. 165–182.

Janda, L./Kopotev, M./Nesset, T. (2020): Constructions, their families and their neighborhoods: the case of *durak durakom* 'a fool times two'. In: Russian Linguistics 44, S. 109–127. https://doi.org/10.1007/s11185-020-09225-y (Stand: 23.3.2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Mel'čuk, I. (1987): Un affix dérivationel et un phrasème syntxique du russe moderne. Essai de description formelle. In: Revue des études slaves 59 (3), S. 631–648.

Mel'čuk, I. (2021). Morphemic and syntactic phrasemes. In: Yearbook of Phraseology 12. Berlin u. a., S. 33–74.

Pavlova, A. (2020): Und ob es Phraseologie ist! In: Engel, Christine/Pohlan, Irina/Walter, Stephan (Hg.): Russland übersetzen / Russia in Translation / Россия в переводе. Festschrift für Birgit Menzel. Berlin, S. 117–132.

Schafroth, E. (2014): Eine Sache des Verstehens: Phraseme als Konstruktionen und ihre Beschreibung in der Lexikographie Französisch/Deutsch. In: Domínguez Vázquez, M. J./Mollica, F./Nied Curcio, M. (Hg.): Zweisprachige Lexikographie im Spannungsfeld zwischen Translation und Didaktik. Berlin u. a., S. 83–112.
http://www.phil-fak.uni-uesseldorf.de/rom4/equipe/schafroth/sch_texte/ (Stand: 23.3.2022).

Šmelëv, D. N. (1977): Sovremennyj russkij jazyk: Leksika. Moskau, S. 327–330.

Švedova, N. Ju. (1960): Očerki po sintaksisu russkoj razgovornoj reči. Moskau, S. 270–279.

Veličko, A. V. (1996): Sintaksičeskaja frazeologija dla russkich i inostrancev: Učebnoe posobie. Moskau.

Ziem, A. (2018): Construction grammar meets phraseology: eine Standortbestimmung. In: Linguistik Online 90 (3), S. 3–19. https://bop.unibe.ch/linguistik-online/article/view/4316/6450 (Stand: 23.3.2022).

Kontaktinformationen

Anna Pavlova
JGU Mainz, FTSK Germersheim
pavloan@uni-mainz.de

Danksagung

Ich bedanke mich bei Dmitrij Dobrovol'skij, Elena Krotova und Katrin Schlund für ihre Unterstützung und Ratschläge. Auch bin ich den Rezensenten herzlich dankbar, deren wertvolle Kritiken zur Qualität dieses Aufsatzes beigetragen haben.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Ralf Plate

WORD FAMILIES IN DIACHRONY

An epoch-spanning structure for the word families of older German

Abstract The 'Word Families in Diachrony' project (WoDia), for which a funding application to the DFG is in preparation, aims to provide a database-driven online research environment that will enable processes of change in the entire historical vocabulary of German to be investigated by focusing on the changes in word families and the individual means of word formation. WoDia will embed the vocabularies of Old High German (OHG), Middle High German (MHG), Old Saxon (OS), and Middle Low German (MLG) in a database, resulting in a word-family structure for High and Low German from the beginnings up to the 15th century (for High German) and up to the 17th century (for Low German). The basis of the vocabulary is provided by reference dictionaries of the four historical varieties, whereas the word families' historical structure is based on the word-family dictionary of OHG by Jochen Splett (1992). Each lemma in the database will be assigned, where appropriate, to a word family. The individual word-formation elements and the word-formation hierarchy will be mapped in a structural formula. The etymologically corresponding lemmas and word families of the different periods/varieties of older German will be linked so that an analysis across the varieties will also be possible. The annotations of word families in the database (e. g., relating to word structure) will be supplemented by linking their lemmas to the online dictionaries and to the reference corpora of Old German (OS and OHG), MHG, and MLG.

Keywords Older German (OHG, MHG, OS, MLG); word family database; historical word formation of German

Introduction

The project presented here will address several desiderata of historical lexicography and historical word-formation research in German:

- The linking of the vocabularies of the older stages of High and Low German (OHG, MHG, OS, MLG), as documented in the reference dictionaries of these varieties, so that each word of each of the varieties is linked to its etymological counterparts in the others (insofar as they are extant).
- The morphological analysis of these vocabularies' word formation, so that for each word, its word-formation structure is recorded in a hierarchical formula.
- The structuring of these vocabularies into word families and the establishment of an overarching word-family structure, in which the individual vocabularies of the four varieties are brought together.

The results of the project will be made available online in WoDia (Word Families in Diachrony), a database-driven research environment with a user-friendly front end for research and teaching.

The project is able to draw on extensive preliminary work and resources, which are described under 1. The basis of the data and the necessary stages of work on the project are presented under 2.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

1. Preliminary work and resources

1.1 Jochen Splett's word-family dictionaries of OHG (Splett 1993) and of contemporary German (Splett 2009)

The reference dictionaries for the above-mentioned varieties of older German are based on single words and do not contain any systematic morphological information about the word formation of the lemmas. However, a comprehensive analysis of word-formation morphology has been carried out by Jochen Splett in two word-family dictionaries: on OHG (1992, 3 vols) and on contemporary German (2009, 18 vols). The particular distinction of these dictionaries, aside from the division of lexemes into word families, is that they offer a hierarchical (bracketed) structural formula for each lemma of a word family, which indicates the constituent structure and the part of speech of the stems. Cf. the following examples from the dictionary of contemporary language (vol. 1, pp. 416f.; cf. also Fig. 1):

- *Bau-er*: (wV)sS – To be read as: substantival derivative from the verbal stem BAU with the suffix *-er* (w = root / stem; V = verb; s = suffix; S = substantive).
- *Acker-bau-er*: (wS) ((wV)sS) – Composite with BAU-ER as the basic component and the stem ACKER as the determinative element.
- *Boots-bau-er*: ((wS) ((wV)S))sS – substantival derivation from the compound noun BOOTS-BAU with the suffix *-er*; *-bau* is again a substantival conversion of the verbal stem BAU.
- *Er-bau-er*: (p(wV))sS – noun derivation from the prefixed verb stem ER-BAU with the suffix *-er* (p = prefix).

The ambiguity arising from double or multiple motivations is systematically taken into account by specifying alternative structural formulae, e. g., in *baukünstlerisch* adj. (vol. 1, p. 414) with three formulae for the three possible analyses (*baukünstlerisch* or *bau-künstlerisch* as immediate constituents and again in the first case *bau-künstler* or *baukunst-ler* as immediate constituents of the first element).

In particular, the structural formulae, with their formalized description of the steps and word-formation elements on which the formations are based, offer a wide range of insights for questions of historical word formation theory. However, since Splett's dictionaries are not yet digitally accessible, their usefulness for such questions has so far been significantly restricted.

1.2 Thomas Klein's semi-automatic word-formation analysis and determination of word family stems

Apart from OHG, the material basis for comprehensive studies of historical word formation, the development and restructuring of vocabulary, and the means of word formation in general is still lacking. This needs to be developed through a morphological analysis of word formation for the historical vocabularies as in Splett's dictionaries. Thomas Klein has taken the initiative in addressing this desideratum, by presenting a methodologically convincing approach to a semi-automatic morphological analysis of word formation and word family classification, in his case, for Middle High German (most recently, Klein 2018). Klein's basic idea consists in the automatic segmentation of lemmas into affixes and stems and the

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

automatic tracing back of the stems to that stem variant which appears in the root (core) word of their word family, the word family stem (WFSt), i. e., the form without OHG *i*-umlaut, ablaut, grammatical change, gemination, Germanic *i*- and *a*-umlaut, etc. In this way, for example, the stems (stem variants) GOLT and GÜLT are isolated from the MHG lexemes *unvergolten* (part. adj.) and *gültic* (adj.), by separating the affixes; they are then traced back to the WFSt GELT. These two and all other words of the MHG word family GELTEN then (ideally) appear under the WFSt GELT.

1.3 Testing of the transepochal word family linking in the ZHistLex Project

Whereas Klein (2018) focuses on the task of automatically determining the MHG WFSt, the further step of semi-automatically mapping the MHG vocabulary to the OHG word families was tested in a pilot project at the University of Frankfurt/Main (a part-project in the preparation of an eHumanties Centre for Historical Lexicography, ZHistLex, funded by the BMBF 2016–2019; cf. Plate 2020).

Exact OHG-MHG word equivalents exist for only one sixth (17%) of the MHG vocabulary; however, for the semi-automatic assignment of the remaining MHG vocabulary to OHG word families, the linking of their stems resulting from the linking of word equivalents can be used, as indicated in the above example (1.2).

Because MHG *unvergolten* was linked to earlier OHG *unfirgoltan* and thus also to its word family GELTAN, its WFSt GELT is also linked to the OHG word family GELTAN. The established correspondences of the type 'MHG WFSt GELT = OHG word family GELTAN' can now be used to assign MHG words automatically to an OHG word family without an OHG precursor if their WFSt (or in the case of compounds, at least one of the word family stems) is already recorded and assigned. Thus, the MHG lexeme *bete-gültic* 'liable to tax' (for which, as well as for its adjectival second component *gültic*, no OHG antecedent is attested) can be automatically assigned to the OHG word family GELTAN on the basis of the MHG WFSt GELT that had already been determined. The same applies to the first component *bete* with the WFSt *bit* and the OHG word family BITTEN. – The automatic assignment gives rise to lists of suggestions, which have to be carefully examined individually.

1.4 Online dictionaries and lemma lists

Lemma lists are a prerequisite for processing the vocabularies of the varieties of older German. These are extracted from the digitized reference dictionaries. In the case of Old High German and Middle High German, the dictionaries and lemma lists are already available digitally.

Reference dictionaries for Old High German are the *Althochdeutsches Wörterbuch* of the Leipzig Academy of Sciences (AWB), which is still being compiled and is available online in the Trier Dictionary Network, and, until the AWB is complete, the *Althochdeutsches Wörterbuch* by Jochen Splett (1992 = AWB^{Spl}), which covers the entire vocabulary of Old High German (approx. 28,500 words). It has been digitized at the University of Frankfurt and will also be published online in the Trier Dictionary Network as part of the project. A sample of the online publication in preparation is shown in Figure 1 on the following page; in the left column the list of word families and lemmas is shown, in the centre column the dictionary,

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

with links to the Leipzig *Althochdeutsches Wörterbuch*, in the right column the results of a search on the structural formulae.

The reference dictionaries for Middle High German are the *Mittelhochdeutsches Wörterbuch* (MWB) of the Göttingen and Mainz Academies of Sciences and Humanities, which is still being compiled and is also available as part of an online offering (MWB Online), and until the MWB is completed, its predecessors, the 19th-century dictionaries (BMZ, Lexer), which are accessible online in the Trier Dictionary Network. MWB Online also contains a complete lemma list of the precursor dictionaries and the MWB (approx. 90,000 words). The period for the sources of the older dictionaries extends into the 15th century, that of the MWB only up to 1350; it is estimated that the MWB will process around 50,400 words. The more comprehensive lemma list of the older dictionaries therefore provides a bridge to older Early New High German.

The *Altsächsisches Handwörterbuch* (ASWB) by Heinrich Tiefenbach (2010) and the *Mittelniederdeutsches Handwörterbuch* (MNWB) by Lasch/Borchling, which will shortly be completed, are not yet available online; they are to be digitized and published online in the Trier Dictionary Network as part of the project. The lemma list of the ASWB comprises around 6,900 words, that of the MNWB around 80,000 words.

1.5 Reference corpora

The reference corpora of the DeutschDiachronDigital initiative for Old German (ReA, OHG and OS), MHG (ReM) and MLG (ReN) cover approximately the same period as the reference dictionaries. ReM and ReN provide material structured according to time, space, and to some extent, text types. All three corpora are grammatically annotated (PoS, morphology) and lemmatized. They can be searched at the levels of tokens, annotations, and lemmas, but as with the dictionaries' online offerings, they are not interlinked. Linking with the reference dictionaries is performed only in the case of ReM (MWB, Lexer).

The vocabulary recorded in the reference corpora is (naturally) considerably smaller than that of the reference dictionaries: the ReA contains the complete textual tradition of OHG and OS but not the extensive tradition of glosses; it comprises 4,100 words for Old Saxon and 10,900 words for Old High German. The corresponding OHG and OS lemmas are not interlinked. ReM contains about 22,300 words, ReN about 17,000.

The linking of WoDia with the reference corpora enables the interconnection and joint use of the corpora, and it opens up an annotated textual evidence base for WoDia.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

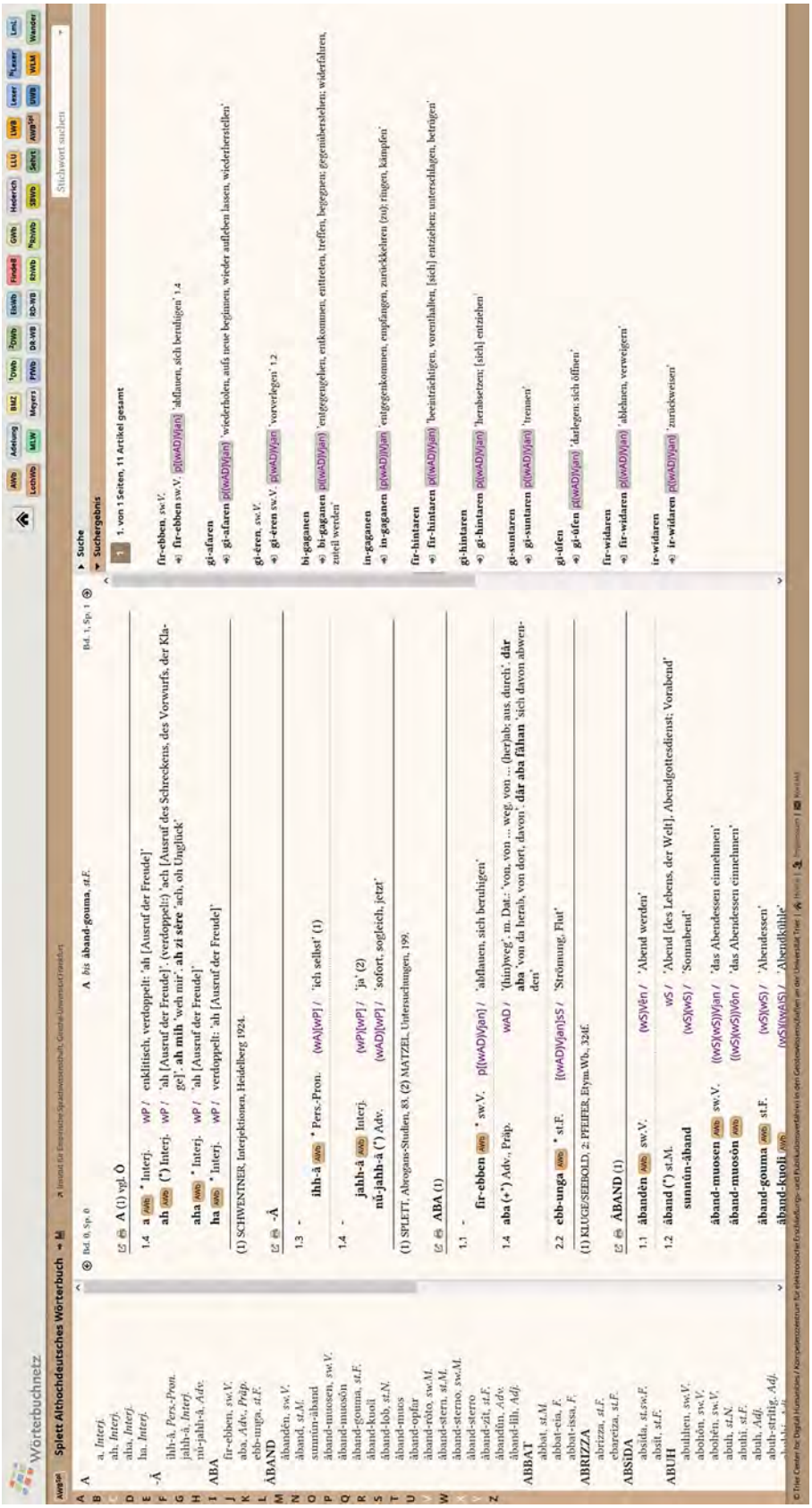


Fig. 1: Jochen Splett, *Althochdeutsches Wörterbuch* (1992), example page of the online publication (in preparation)

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

2. Basic data and work stages of the project

2.1 Basic ata

The project's basic data are the lemma lists of the reference dictionaries, for OHG additionally the word family structure and the structural formulae of AWB^{Spl}. Figure 2 summarizes the scope of the lemma lists included in WoDia (cf. 1.4 above) and the lemma lists of the reference corpora to be linked with them (cf. 1.5 above):

Variety	Dictionary	Reference Corpus
OS	6,900 (ASWB)	4,100 (ReA)
OHG	28,500 (AWB^{Spl})	10,900 (ReA)
MHG	90,000 (Lexer, to 15th century) including 50,400* (MWB to 1350)	22,300 (ReM)
MLG	80,000* (MNWB)	17,000 (ReN)

Fig. 2: Sizes (number of lemmas) and sources (dictionaries) of the vocabularies in WoDia; sizes of the vocabularies in the corresponding reference corpora; *= estimated number of lemmas when complete

2.2 Stages of work

In the following, the stages of work necessary for setting up WoDia are presented in a simplified version by comparison with the detailed project description of the DFG application.

2.2.1 Digitization of dictionaries and production of lemma lists

The two concise dictionaries ASWB and MNWB will be digitized and published in the Trier Dictionary Network. The lemma lists will then be extracted from the digitized data.

The lemma lists of OHG and MHG are already available. In the case of MHG, we shall mark that part of the vocabulary of the total lemma list from MWB + Lexer that is only attested after 1350 and thus falls within the scope of sources for the Early New High German dictionary that is currently being compiled. This is already apparent for the section so far processed in the MWB (i. e., lemmas of the complete list that, for reasons of dating, have not been included in the MWB); for the remaining parts, marking can be performed semi-automatically with the help of a list of the more recent sources in Lexer. Consequently, the vocabulary of older Early New High German is already incorporated in WoDia, and some groundwork for the inclusion of the Early New High German Dictionary's lemma list has already been performed. However, for reasons of scope and lack of preliminary work, this expansion of WoDia will have to be deferred for the time being.

2.2.2 Linking of the lemma lists

A prerequisite for the assignment to word families across linguistic epochs (cf. 2.2.5 below) is the linking of the lemmas that correspond etymologically. The linking of the OHG and MHG lemma lists has already been performed semi-automatically by Thomas Klein with the

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

help of a suitable script (cf. Klein 2018, pp. 13–16). For the linking of the OS and MLG lists, similar scripts will be written.

2.2.3 Segmentation of the llemmas

For the morphological word-formation analysis, the lemmas are segmented into prefixes, stems, and suffixes on the basis of affix lists. OHG has already been analysed in AWB[Spl]; for MHG, the analysis has been performed semi-automatically by means of a script by Klein 2018 (see above under 1.2). For OS and MLG, similar lists will be compiled and scripts developed.

2.2.4 Determination of word family stems

The stems obtained in 2.2.3 for OS and MLG are traced back to their WFSts by eliminating variants (umlaut, ablaut, grammatical change, gemination, etc.). For this, scripts similar to those used by Klein for MHG are needed. The rules for such elimination contained in these scripts must be developed anew for OS and MLG.

2.2.5 Assignment of word family stems to word families

Suggestions for the assignment to word families are generated for the WFSts determined in 2.2.4, using the direct correspondences already linked in 2.2.2 and with the help of the ZHistLex script for MHG (cf. above under 1.3) as well as the scripts that will be adapted for OS and MLG. OS WFSts will be linked to OHG word families and MLG WFSts to MHG word families.

Owing to the identical spelling of many WFSts, the suggestions are not only 1:1 correspondences but mostly multiple in nature so that these have to be examined individually; cf. for example Klein's (2018, p. 26) example of the homographic WFSt WINT, which belongs to the following word families: *winden* stV. (3a) 'to wind' – *wint* stM. 'wind' – *wint* stM. 'greyhound (Wendish dog)'.

If it is not possible to assign the OS and MHG WFSts to an OHG word family and the MLG WFSts to an MHG word family, equivalents are sought in the other varieties; if no word family can be assigned, the lemmas in question remain as individual words.

2.2.6 Analysis of the constituent structure / structural formulae

A central function of the research environment is the representation of the word families' inner structure, i. e., the stages of word formation that precede the composites and derivations, as expressed in Splett's work and elsewhere by means of bracketed formulas. As in Splett's work, the part of speech of the stems needs to be indicated in the formula. For a minority of cases in the vocabulary of MHG, OS, and MLG, where direct equivalents are present in OHG, the structural formulae of Splett can be transferred; for all other lemmas, the structural formulae have to be elaborated anew. This is done as far as possible through semi-automated processes, in which proposals are generated and then reviewed. A prerequisite for this is the segmentation of lemmas into stems and affixes (2.2.3) and the assignment of stems to word families (2.2.5). For example, a script for a complex lemma can search for freely occurring equivalents of its constituents within the same word family and adopt the specification for their part-of-speech or even a structural formula for the constituent that already exists through direct linking. Regularities that can be used for the semi-auto-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

matic analysis are already observed in the work on 2.2.5 and will be converted into corresponding scripts for the analysis of the constituent structure. In the processing, OS and MHG precede MLG, so that the already existing OS and MHG structural formulae can be transferred for the direct equivalents in MLG.

2.2.7 Setting up a web application

For the overall data management, a central database is created as a two-tier web application, with a component for data collection and modelling (back end) and one for publication and use (front end). The database is connected to the online dictionary resources and the reference corpora via open interfaces and web services. Basic functions of the front end are the display of word families (word family list, display of a specific word family, selection option for the varieties, etc.) and search functions for lemmas, word families, word formation structures, word formation elements, etc.

3. Conclusions

The establishment of WoDia marks significant progress in the development of a digital research infrastructure for German language history. This is done by bundling and supplementing existing digital resources (dictionaries), by introducing a new dimension of description and investigation (word families) across epochs and varieties, by the consistent use of semi-automatic procedures for the sophisticated analysis of extensive language data (vocabularies), and finally by making the results of the work available in a user-friendly web application for research and teaching, with links to the dictionary and corpus resources available online via web services.

References

ASWB = Tiefenbach, H. (2010): Altsächsisches Handwörterbuch. Berlin/New York.

AWB = Althochdeutsches Wörterbuch (1952 ff.): Auf Grund der von Elias v. Steinmeyer hinterlassenen Sammlungen im Auftrag der Sächsischen Akademie der Wissenschaften zu Leipzig. http://awb.saw-leipzig.de/cgi/WBNetz/wbgui_py?sigle=AWB (last access: 24-05-2022).

BMZ = Benecke, G. F./Müller, W./Zarncke, F. (1854–1866): Mittelhochdeutsches Wörterbuch. Leipzig. Digitalisierte Fassung im Wörterbuchnetz des Trier Center for Digital Humanities, Version 01/21. https://www.woerterbuchnetz.de/BMZ (last access: 24-05-2022).

DeutschDiachronDigital [Reference corpora on German language history]. https://www.deutschdiachrondigital.de/ (last access: 24-05-2022).

FWB = Frühneuhochdeutsches Wörterbuch. Bearbeitet von Oskar Reichmann et al. (1986 ff.): Berlin/New York. https://fwb-online.de/ (last access: 24-05-2022).

Klein, T. (2018): Mittelhochdeutsche Wortfamilien: Ermittlung und Perspektiven. In: Zeitschrift für Wortbildung/Journal of Word Formation 2/1, pp. 11–31. DOI: https://doi.org/10.3726/zwjw.2018.01.01 (last access: 24-05-2022).

Lexer = Mittelhochdeutsches Handwörterbuch von Matthias Lexer (18721878). Leipzig. Digitised version in the Dictionary Network of the Trier Centre for Digital Humanities, Version 01/21. https://www.woerterbuchnetz.de/Lexer (last access: 24-05-2022).

MNWB = Lasch, A./Borchling, C. (1956 ff.): Mittelniederdeutsches Handwörterbuch. Begr. von Agathe Lasch und Conrad Borchling. Kiel/Hamburg.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

MWB = Mittelhochdeutsches Wörterbuch (2006 ff.): Im Auftrag der Akademie der Wissenschaften und der Literatur Mainz und der Akademie der Wissenschaften zu Göttingen. Stuttgart. – MWB Online: http://www.mhdwb-online.de/index.html (last access: 24-05-2022).

Plate, R. (2020): Computergestützte Etablierung epochenübergreifender Wortfamilienstrukturen [ZHistLex-Teilprojekt]. Final report. https://zhistlex.de/ziele/wortfamilien/ (last access: 24-05-2022).

ReA (Reference corpus of Old Saxon and Old High German). https://www.deutschdiachrondigital.de/rea (last access: 24-05-2022)

ReM (Reference corpus of Middle High German): https://www.linguistics.rub.de/rem/ (last access: 24-05-2022)

ReN (Reference corpus of Middle Low German): https://www.slm.uni-hamburg.de/ren/ (last access: 24-05-2022)

Splett, J. (1993): Althochdeutsches Wörterbuch. Analyse der Wortfamilienstrukturen des Althochdeutschen. Zugleich Grundlegung einer zukünftigen Strukturgeschichte des deutschen Wortschatzes. 3 vols. Berlin/New York.

Splett, J. (2009): Deutsches Wortfamilienwörterbuch. Analyse der Wortfamilienstrukturen der deutschen Gegenwartssprache. Zugleich Grundlegung einer zukünftigen Strukturgeschichte des deutschen Wortschatzes. 18 vols. Berlin/New York.

Trier Dictionary Network: https://woerterbuchnetz.de (last access: 24-05-2022).

Contact Information

Ralf Plate
Akademie der Wissenschaften und der Literatur | Mainz, Mittelhochdeutsches Wörterbuch, Arbeitsstelle an der Universität Trier
plate@uni-trier.de

Acknowledgements

This article is based on the project description prepared jointly by Jost Gippert (Frankfurt), Sarah Ihden, and Ingrid Schröder (Hamburg), Thomas Burch (Trier), and the author. I should also like to thank Claudia Wich-Reif (Bonn) for her valuable comments. David Yeandle deserves my special thanks for translating the text from German into English.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Kyriaki Salveridou/Zoe Gavriilidou

COMPILATION OF AN ANCIENT GREEK – MODERN GREEK ONLINE THESAURUS FOR TEACHING PURPOSES: MICROSTRUCTURE AND MACROSTRUCTURE

Abstract To effectively design online tools and develop sophisticated programs, for the teaching of Ancient Greek language, there is a clear need for lexical resources that provide semantic links with Modern Greek. This paper proposes a microstructure for an online Ancient Greek to Modern Greek thesaurus (AMGthes) that serves educational purposes. The terms of this bilingual thesaurus have been selected from reference Ancient Greek texts, taught and studied during lower and upper secondary education in Greece. The main objective here is to build a semantic map that helps students find relevant and semantically related terms (synonyms and antonyms) in Ancient Greek, and then provide a rich set of suitable translations and definitions in Modern Greek. Designed to be an online resource, the thesaurus is being developed using web technologies, and thus will be available to every school and university student that pursues a degree in digital humanities.

Keywords Online thesaurus; bilingual thesaurus; Ancient Greek; pedagogical lexicography

1. Introduction

The conceptual approach constitutes the core of what we call conceptually organized lexica, onomasiological dictionaries or thesauri, all of which are based on ontologies. In such ontologies the world is structured in hierarchically organized thematic fields. Every field represents a node in the hierarchy that groups together related linguistic signs. These signs are essentially terms linked to each other with semantic relations such as hyponymy, hypernymy, synonymy etc. The goal of lexica that host such structures is to support users in understanding the semantic value of these relations and enable them to access a rich collection of terms. This in turn will help them identify the words that exactly capture their thoughts, upgrading their vocabulary in the process. Furthermore, there is a clear need for tools that serve as conceptual dictionaries and thesauri. The availability of such tools varies significantly according to the language, scope and targeted function (i. e. educational purposes). Specifically, Ancient Greek lacks essential linguistic digital resources. The question arises: How can we use modern technologies to enrich conceptually organized dictionaries?

This paper discusses the design features, the structure and the implementation of an online bilingual thesaurus for Ancient – Modern Greek (AMGthes). The goal is to deliver a web resource that is oriented towards supporting educational functions. In greater detail, the proposed thesaurus aims to aid the better understanding, accurate interpretation and in-depth analysis of Ancient Greek texts, by both high school and university students (humanities).

The paper is organized in the following sections: Section 2 offers the motivation for the creation of the AMGthes along with capturing relevant research. Section 3 provides the basic concepts and the context of the theoretical background. Section 4 describes the methodology that was followed and the process of selecting and filtering the corpus. This section

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

also analyzes the macrostructure and the microstructure of the thesaurus, explaining the deciding features and components respectively. Section 5 discusses the details for the design and development of the online resource and finally section 6 presents our conclusions and the next steps planned.

2. Motivation for the creation of the AMGthes

The value of dictionaries and the impact of their use during learning, is well established by relevant research (Atkins 1998; Dziemianko 2012a; Gavriilidou/Konstantinidou 2021).

Bilingual thesauri, on the other hand, are a valuable tool in the hands of students and the whole educational context can greatly benefit from such lexical resources, especially when their continuous improvement both in terms of accuracy and content is consistently worked upon. Modern thesauri and dictionaries are powered by advanced implementations and can now be accessed as online platforms that effectively support the learning context of any scale or complexity. Making academic versions of such tools more available to scholars and better adapted to their needs, will help address fragmentation within the pedagogical discipline and fuel the research of new solutions, thus allowing a methodological shift (Bizzoni et al. 2014; Berti et al. 2016).

While significant steps have been made, the availability of digital resources capable of supporting the teaching of Ancient Greek are still limited and not actively maintained. Modern versions (web-based, online) for established tools and platforms are not being developed, since most are treated as legacy systems with rare or no updates. More importantly, their content is not often curated with scheduled contributions from the community, thus reducing the time window of their validity. In order to gain momentum and increase their user base, digital lexical tools for the Ancient Greek language have to be systematically enriched with new knowledge and upgraded to utilize modern technologies and access patterns (i. e. web or mobile). Attempting to address this need, our research delivers an online bilingual thesaurus that hosts the vocabulary wealth from the corpora that currently comprises the syllabus for teaching the Ancient Greek language at Greek schools. AMGthes adopts a widely known semantic classification scheme, to power its taxonomy, and employs no pivoting language, to facilitate the translation of the selected terms and making it an integral part of their learning process.

3. Theoretical background

The design of the proposed bilingual thesaurus was based on two well-established and widely used thesauri with significant value and impact on many modern linguistic resources: the Roget's Thesaurus (1911) for the English Language and the Antilexicon (1962) by Vostantzoglou for the Greek Language.

In 1852 Peter Mark Roget's Thesaurus, known as the Roget's Thesaurus of English words and phrases, occupied a prominent position in English Lexicography. In this thesaurus, words and terms are not presented alphabetically, but instead grouped together according to the semantic relations that bind them (Hüllen 2004). Roget's Thesaurus comprises one of the largest collections of semantically organized terms, and Roget's name has become a synonym for thesaurus (Jarmasz 2012).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Adopting Roget's classification theme for organizing terms and concepts, the Antilexicon by Vostantzoglou (1862) is a reference Greek lexical resource and a typical example of a similarly structured thesaurus. The Antilexicon is considered a versatile linguistic tool that facilitates effective learning of Greek language. A key feature of this thesaurus is that Vostantzoglou extensively analyzes each term, managing to populate a rich profile with many correlations and semantic links (Trapalis et al. 2005). At its core, linguistic material is grouped into categories of semantically related words, distinguished by individual properties. The Antilexicon comprises a total of 1.500 such categories and interpretations are provided along with both grammar and syntax, while Roget's thesaurus offers no such type of information.

In order to compile the AMGthes, we also focused on tools that capture and analyze the concepts and semantics of words, such as WordNet. This lexical database included English nouns, verbs, adjectives and pronouns, all semantically organized in sets of synonyms, named synsets (Miller/Fellbaum 2007). From relevant previous research, the Ancient Greek WordNet (AGWN) is the first attempt to create the WordNet version for the Ancient Greek language. This version followed the main paradigm of Princeton WordNet (PWN), while adopting features from other WordNets such as the Italian and the Latin versions. This approach resulted in a thesaurus where users can access the synsets' equivalents for each different language (Bizzoni et al. 2014).

AGWN started with evaluating and analyzing bilingual dictionaries to bootstrap Greek – English pairs. As a first step, Greek words were linked to PWN synsets that included the respective English translations. An issue that emerged from using this approach was the invalid overpopulation of synsets and semantic relations, due to the multifaceted nature of English homonymy and polysemy. Furthermore, AGWN is not tailored to effectively serve the learning requirements and modern needs of Greek students. This motivated us to study and create a new online thesaurus, whose design and development are discussed and analyzed in the following sections.

4. Methodology

According to Gouws/Prinsloo (2010) the lexicographic protocol or dictionary plan is a workflow of well-defined tasks, which includes two main components: the organization plan and the dictionary conceptualization plan. Wiegand (1998, p. 151) describes the dictionary conceptualization plan as a process of five distinct phases: the general preparation phase, the material acquisition phase, the material preparation phase, the material processing phase and the publishing preparation phase. Adapting this workflow, Klosa (2013) offers a computer-lexicographical process for the development of online dictionaries. Her research examines a plan that includes: the phase of preparation, the phase of data acquisition, the phase of computerization, the phase of data processing, the phase of data analysis and the phase of preparation for online release.

Considering the above protocols, the compilation of AMGthes follows Klosa's process phases, while also adopting the planning for continuous maintenance and improvement of both the dictionary and its online implementation. All the decisions relevant to the design and features of AMGthes macrostructure and microstructure were made in accordance with this protocol and its methodological principles. Benefiting from modern lexicographical practices and methods, AMGthes is based on a workflow that determines how best to plan, develop, control and deliver its content in an online format. Currently, its progress reached the data

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

preparation phase for online release, with an alpha version already available and pilot tested. Apart from finalizing and testing its implementation, our work now focuses on conducting multiple sprints of data processing and analysis for the expansion of its vocabulary with new terms.

4.1 Selecting the entries

The first stage of compiling the AMGthes vocabulary included tasks that focused on extracting, collecting and capturing the initial set of terms, from the Ancient Greek texts that comprise the teaching curriculum of Greek middle and high school courses. These tasks span across the phases of data preparation, acquisition, and computerization.

As part of the preparation phase, a set of school courses were selected, forming the official syllabus for learning the Ancient Greek language. Each course includes the analysis and interpretation of texts, from the work of famous Ancient Greek philosophers, poets, orators and historians. In terms of data acquisition, all the appropriate texts were accessed and retrieved from Photodentro – the Greek National Aggregator of Educational Content.[1] A corpus was created from the texts of the appropriate courses: i) For all three grades of middle school this includes excerpts from various Ancient Greek authors. ii) For the 1st grade of high school, excerpts from Xenophon's Hellenica and Thucydides' History of the Peloponnesian War. iii) For the 2nd grade of high school, the Speech of Lysias for Manditheos. iv) For the 3rd grade of high school excerpts from the book Philosophical Reason, including among others Aristotle's Politics and Nicomachean Ethics, Protagoras and Plato's Republic. Odyssey, Iliad, Antigone and other subjects, while being part of the curriculum, they are taught as translated texts and thus were not included.

Completing the computerization phase, the above corpora was processed, organized and stored into multiple text files, properly formatted to act as the input of AntConc,[2] a text analysis software tool. A starting set of 5.566 unique words was retrieved from the text analysis, acting also as the first task of the data analysis phase. These words were filtered and reviewed for further refinement, detecting misspelled duplicates and other minor mistakes. The resulting list was then studied, identifying the lexical category of each word and bringing it to its basic or citation form. In open lexical categories, i) the citation form of a noun is the singular nominative, ii) the citation form of a verb is the first person singular, present tense in active voice, and iii) the citation form of an adjective is masculine singular nominative. Any category other than verbs, nouns, adjectives and pronouns were filtered out, creating a final set of 2.052 candidate words for the AMGthes vocabulary. These steps can be classified as tasks for both data processing and data preparing.

As an example of how the final set of terms is evaluated and studied, the term "εὐνομία" / "favor", found in one of the Ancient Greek texts, is coupled with the term "νόμος" / "law" as its hypernym, and then labeled as a member of "ήθος" / "ethos" semantic class. Following the same process, the term "αὐτάρκεια" / "self-sufficiency" is coupled with its hypernym "ἐπάρκεια" / "sufficiency" and labeled as a member of "ποσότητα" / "quantity" semantic class. Processing terms this way, gradually builds the semantic context for a growing num-

[1] Photodentro Greek National Aggregator of Educational Content (http://photodentro.edu.gr/).

[2] AntConc Text Analysis Software (https://www.laurenceanthony.net/software/antconc/).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

ber of concepts, grouping together relevant terms and phrases as part of their semantic grid. Following such relations of hyponymy and hypernymy allows the above process to step up from the semantic level of the initial terms and focus on concepts that consistently participate as thematic pillars of the Ancient Greek texts.

The semantic analysis of terms, for an Ancient Greek to Modern Greek bilingual thesauri, requires the cross referencing and extensive studying of relevant thesauri and dictionaries. Currently, there is a sufficient number of curated Ancient to Modern Greek dictionaries, able to facilitate the accurate translation and interpretation of terms. On the other hand, only a small number of Ancient Greek thesauri are available, all quite old and most having received few to no updates. As a result, conducting a semantic analysis for a large volume of terms, for an Ancient to Modern Greek bilingual thesauri, is an ambitious feat that requires time, dedication and making the most out of the above resources. This work's goal is to study the extracted set of unique terms and produce a piloting set of concepts that will be consistently expanded and maintained. These are concepts that hold significant semantic and educational value, supporting the teaching of the respective Ancient Greek texts.

4.2 Macrostructure

Defining the semantic structure of the AMGthes was a result of studying other well established bilingual thesauri. The purpose of this task was to understand and adopt semantic relations from both old and modern lexical resources with similar goals and targeted functions. More specifically, the classification scheme of three thesauri was studied in detail: i) the Roget's Thesaurus, ii) Antilexicon by Vostantzoglou, and iii) the Cambridge French – English Thesaurus by Marie Noelle Lamy (1997). The decided macrostructure is based on a taxonomy that is essentially a combination of the above, primarily adopting the semantic classes of Vostantzoglou. The reason for this alignment is that Vostantzoglou is an older lexical resource that analyzes terms which are conceptually more relevant to the semantic context of the Ancient Greek texts. In contrast to more modern thesauri, Vostantzoglou's vocabulary is more compatible with Ancient Greek concepts and as such his taxonomy offers classification with improved semantic accuracy.

In terms of typological features, AMGthes can be categorized as an online, bilingual, special, conceptual dictionary. This classification denotes that the provided information is organized, based on semantic criteria, and helps users find language signs for specific concepts (Atkins/Rundell 2008 p. 4). Aligned with Antilexicon, AMGthes too functions as a hierarchical hypernymic taxonomy of concepts that sorts its terms based on their semantic class. At its top level, the semantic taxonomy of AMGThes is divided into General and Special concepts. General concepts are then subdivided in Abstract and Specific, while Special into Tangible and Intangible. Each subclass is further analyzed as follows (Fig. 1):

- [General][Abstract]: Existence, Relation, Unity, Order, Quantity, Number, Time
- [General][Specific]: Space, Dimension, Shape, Energy, Motion
- [Special][Tangible]: Inorganic Matter, Organic Matter
- [Special][Intangible]: Mind, Will, Action, Values, Emotion, Ethos, Religion

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

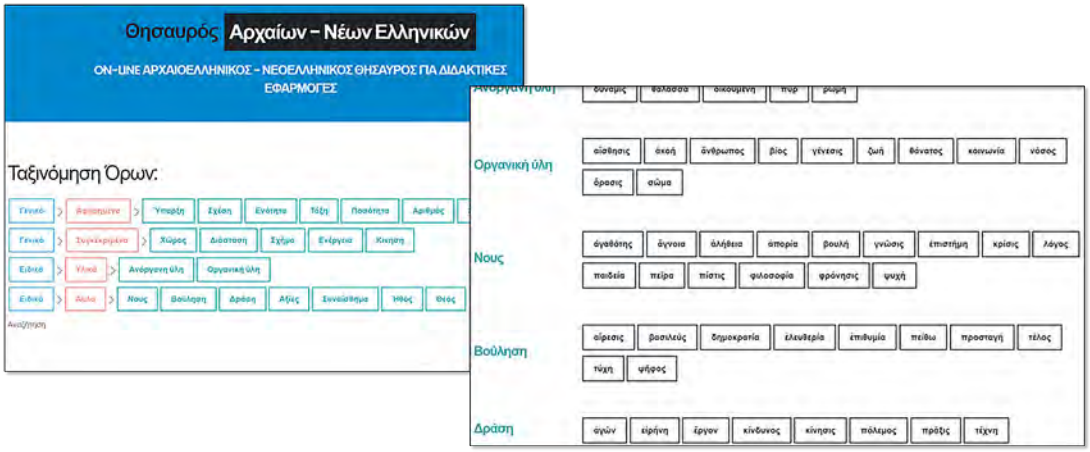


Fig. 1: Macrostructure Semantic Classes – Term Classification Catalog

Being an online dynamic resource, AMGthes can also render an alphabetically ordered list of its terms. Its users can navigate and view the terms included in each of the above semantic classes. Furthermore, both of its currently available catalogs include a search field, allowing the dynamic filtering of their content (Fig. 1). It is worth mentioning that AMGThes data representation and data persistence (the way information is stored and handled) is future proofing its ability to expand and use new semantic classes. The above taxonomy and macrostructure are stored and structured in a human-readable data format (JSON), and as such they can be easily updated and extended to support new semantics and mechanisms for the optimal organization of the constantly enriched AMGthes vocabulary.

4.3 Microstructure

As mentioned before each term is assigned to a semantic class defined as part of the macrostructure taxonomy. Combining information from the decided reference lexical resources, each concept is extensively studied and researched, to identify and map the semantic relations of synonymy, antonymy, hypernymy and frozen expressions. These are the core components for building the semantic profile of each term and thus creating the grid of semantic relations of AMGthes.

The AMGthes microstructure is organized into modules, each profiling a general concept (a thesaurus lemma) by providing synonyms and related words in Ancient and Modern Greek. The large majority of concepts are nouns. The relation of antonymy is also addressed and studied as a separate concept. Antonym concepts are coupled together and presented side by side as linked modules. Inheriting its formalizations from the three reference thesauri, AMGthes renders its modules as online content, using the following template and formatting rules (Fig. 2):

- As a title at the top of each module the reference number (code) and the lemma (the concept) are displayed in bold text. Next to them, the semantic class of the concept is displayed between brackets. Every module is displayed in one column of the page. In case an antonym concept has been studied and profiled, its module will be displayed at the opposing side as a second column.
- A module contains different sections for each lexical category of available synonyms and related words. This means that all the terms (Ancient or Modern Greek) inside a section

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

share the same lexical category. Each section starts with an abbreviation of the lexical category: "Ουσ." for nouns, "Ρ." for verbs, "Επιθ." for adjectives and "Επιρ." for pronouns. Each section lists every Ancient Greek synonym or related word of the main concept in a separate line. The first synonym or related word is displayed right next to the section's abbreviation for its lexical category. A module may also feature an extra section for phrases, using the abbreviation "Φρ.". Each phrase is formatted in bold text, followed by an optional reference source and relevant Modern Greek phrases.

– Every Ancient Greek synonym or related word is formatted in bold text. It is followed by an optional definition in Modern Greek between square brackets. This optional definition is followed by a set of Modern Greek synonyms or related words, separated by commas. This part of each section functions as a bridge between Ancient and Modern Greek, providing the semantic links that enable AMGthes to serve as a bilingual thesaurus.

To instantiate the above microstructure for every selected concept and produce its content, both online and printed versions of widely approved dictionaries were studied. This extended task represents the data processing and data analysis phase of AMGthes. The digital version of Lidell Scott dictionary[3] served as the primary resource for Ancient Greek, along with the Stamatakou dictionary.[4] For Modern Greek, the Antilexikon by Vostantzoglou provided the main resource of material and the overall point of reference. The dictionaries by Dimitrakou[5] and Triantafyllidis,[6] along with Babiniotis[7] dictionary of synonyms and antonyms, also supported the data processing in Modern Greek. Proper fusion of semantics and definitions from all the above resources was critical for producing balanced modules, suitable for learning purposes and consumption by students.

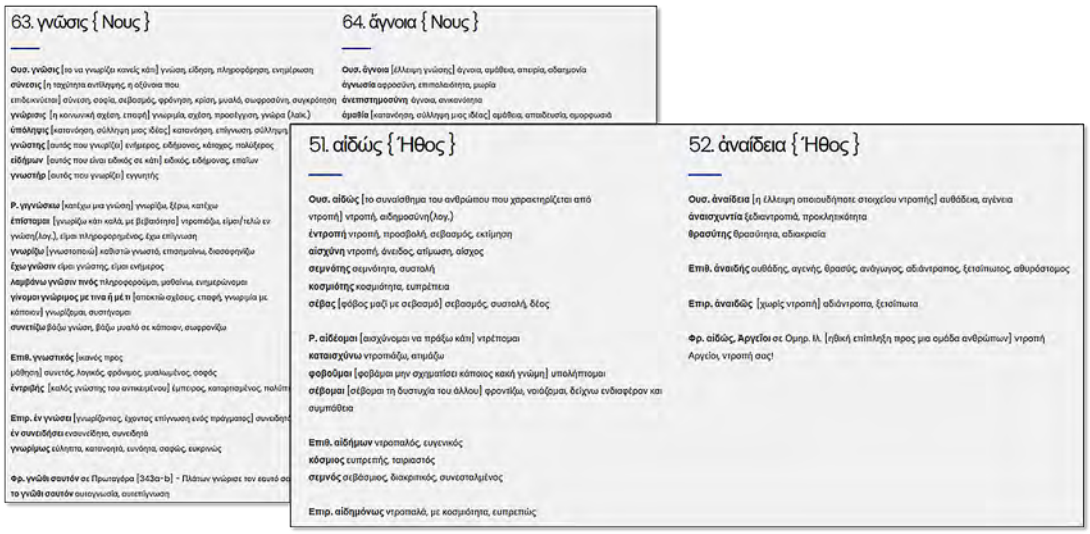


Fig. 2: Microstructure for Concepts "αἰδώς", "γνῶσις" and their antonyms

[3] Liddell Scott online dictionary (https://www.greek-language.gr/digitalResources/ancient_greek/tools/liddell-scott/).

[4] Stamatakos, I. (1972): Dictionary of the Ancient Greek language. Athens.

[5] Dimitrakos D. (1953): Great lexicon of the Greek language. Domi Publications.

[6] Triantafyllidis online dictionary (https://www.greek-language.gr/greekLang/modern_greek/tools/lexica/triantafyllides/).

[7] Babiniotis, G. (2011): Dictionary of synonyms and antonyms of Modern Greek. Lexicology Center.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

5. Implementation of the online thesaurus

Following the description of the thesaurus' core semantic components and the reference term's analysis, this section offers details about the online features and web-based implementation. The online thesaurus is being developed as a dynamic website that utilizes modern technologies for its front-end interfaces, back-end services and data instantiation methods.

More specifically, the online thesaurus offers two views, for the presentation of its terms: i) a sorted list of terms in alphabetical order and ii) a semantic taxonomy of terms based on the decided macrostructure (Vostantzoglou classification scheme). Both interfaces were developed with HTML, CSS and JavaScript technologies, allowing for page designs that are aesthetically appealing and easy to navigate. Each listing provides internal links for quick access into the alphabetical order and the taxonomy classes, along with a search bar that facilitates the filtering of terms.

In terms of back-end implementation, instead of offering a simple static web page, the proposed thesaurus functions as a dynamic resource. Its views/interfaces are powered by web services, developed in JavaScript using the Node.JS platform. These services access and process the files, where the structured data for each term's semantic profile is stored. To properly capture the information for a populated microstructure, our data instantiation employs the widely used JSON format, currently driving most internet data and metadata. These data modeling decisions provide the means for the versatile referencing inside each term's profile, giving direct access to specific semantic relationships and their translations.

The online thesaurus is built with modern technologies that ensure its responsiveness and fast rendering. Both its mechanisms (services) and information architecture, offer the means for future extensions and enriched content. The design theme and navigation mechanisms, ensure efficient browsing and an overall pleasant experience while exploring this online lexical resource.

6. Conclusions

Discussing the design decisions for the AMGthes, this paper highlights its development in terms of content and semantic features. The role of a lexicographic process is studied, describing the phases and compilation tasks of the adopted workflow. Starting with the data acquisition of unique terms, from Ancient Greek texts of the Greek school curriculum, our work gradually prepares, processes, and models the data and semantics of AMGthes. Its macrostructure and microstructure are explained, profiling their alignment with methods and techniques from the employed reference dictionaries and thesauri. Finally, the paper provides information regarding the online implementation and web access pattern. The role of modern technologies is described, enabling the development of a dynamic web portal with data handling mechanisms. By design, its online content will support scheduled updates, improvements and optimizations, offering a consistently enriched resource for students and academics.

AMGthes is built using a pipeline of compilation tasks that effectively fuse knowledge to produce a much-needed lexical resource. Creating and maintaining AMGthes is a challenge and a requirement for bringing the teaching of Ancient Greek language up to speed with modern learning technologies and teaching strategies. The presented work delivers the pilot version of a bilingual thesaurus that trains the students' ability to traverse semantic rela-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

tionships and empowers them to explore, translate and accurately interpret the context of Ancient Greek texts. Upon the completion of AMGthes development and lexical coverage expansion, it will be hosted at the webpage of +MorPhoSe lab of Democritus University of Thrace, free to be accessed by students and any prospect visitors.

References

Atkins, S. B. T. (ed.) (1998): Using dictionaries. Studies of dictionary use by language. Learners and translators. Tübingen.

Atkins, B. S./Rundell, M. (2008): The Oxford guide to practical lexicography. Oxford.

Babiniotis, G. (2011): Dictionary of synonyms and antonyms of Modern Greek. Lexicology Center.

Berti, M./Bizzoni, Y./Boschetti, F./Crane, G./Del Gratta, R./Yousef, T. (2016): Ancient Greek WordNet meets the dynamic lexicon: the example of the fragments of the Greek historians. In: Proceedings of the 8th Global WordNet Conference (GWC), pp. 34–38

Bizzoni, Y./Boschetti, F./Diakoff, H./Del Gratta, R./Monachini, M./Crane, G. (2014): The making of ancient greek wordnet. In: Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), pp. 1140–1147.

Dimitrakos, D. (1953): Great lexicon of the Greek language. Domi Publications.

Dziemianko, A. (2012a): On the use (fulness) of paper and electronic dictionaries'. In: Granger, S./Paquot, M. (eds.): Electronic lexicography. Oxford.

Gavriilidou, Z./Konstantinidou, E. (2021): The design of an explicit and integrated intervention program for pupils aged 10–12 with the aim to promote dictionary culture and strategies. In: Gavriilidou, Z./Mitis, L./Kiosses, S. (eds.): XIX Euralex Proceedings. Vol. 2, pp. 735–745.

Gouws, R. H./Prinsloo, D. J. (2010): Principles and practice of South African lexicography. Stellenbosch.

Hüllen, W. (2004): A history of Roget's thesaurus: origins, development, and design. Oxford.

Jarmasz, M. (2012): Roget' s thesaurus as a lexical resource for natural language processing. arXiv preprint. arXiv:1204.0140.

Klosa, A. (2013): The lexicographical process (with special focus on online dictionaries). In: Hausmann, F. J./Reichmann, O./Wiegand, H. E./Zgusta, L. (eds.): Wörterbücher. Ein internationales Handbuch zur Lexikographie. Supplement Volume: Recent developments with focus on electronic and computational lexicography. (= Handbücher zur Sprach- und Kommunikationswissenschaft (HSK) 5.4). Berlin/Boston, pp. 517–524.

Lamy, M. N. (1997): The Cambridge French-English thesaurus. Cambridge.

Liddell Scott Online Dictionary.
https://www.greek-language.gr/digitalResources/ancient_greek/tools/liddell-scott/.

Miller, G. A./Fellbaum, C. (2007): WordNet then and now. Language resources and evaluation. In: Language Resources and Evaluation 41 (2), pp. 209–214.

Roget, P. M. (1911): Roget's thesaurus of English words and phrases. New York.

Stamatakos, I. (1972): Dictionary of the Ancient Greek language. Athens.

Trapalis, G./Markantonatou, S./Alexopoulou, A./Fotopoulou, A./Maistros, Y. (2005): The Antilexicon by Vostantzoglou and the development of a small scale semantically organized lexical database of Modern Greek. In: 7th International Conference on Greek Linguistics, York, UK, September 8–10.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Triantafyllidis Online Dictionary.
https://www.greek-language.gr/greekLang/modern_greek/tools/lexica/triantafyllides/.

Vostantzoglou, Th. (1962): Antilexikon or onomasticon of Modern Greek language. Athens.

Wiegand, H. E. (1998): Wörterbuchforschung: Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie. Berlin/New York.

Contact information

Kyriaki Salveridou
Democritus University of Thrace
ksalveri@helit.duth.gr

Zoe Gavriilidou
Democritus University of Thrace
zgabriil@helit.duth.gr

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Bilingual Dictionaries

XX EURALEX INTERNATIONAL CONGRESS
DICTIONARIES AND SOCIETY
12-16 July 2022, Mannheim, Germany

Voula Giouli/Anna Vacalopoulou/
Nikos Sidiropoulos/Christina Flouda/Athanasios Doupas/
Gregory Stainhaouer

FROM MYTHOS TO LOGOS: A BILINGUAL THESAURUS TAILORED TO MEET USERS' NEEDS WITHIN THE ECOSYSTEM OF CULTURAL TOURISM

Abstract Thesauri have long been recognized as valuable structured resources aiding Information Retrieval systems. A thesaurus provides a precise and controlled vocabulary which serves to coordinate data indexing and retrieval. The paper presents a bilingual Greek and English specialized thesaurus that is being developed as the backbone of a platform aimed at enhancing and enriching the cultural experiences of visitors in Eastern Macedonia and Thrace, Greece. The cultural component of the intended platform comprises textual data, images of artifacts and living entities (animals and plants in the area), as well as audio and video. The thesaurus covers the domains of Archaeology, Literature, Mythology, and Travel; therefore, it can be viewed as a set of inter-linked thesauri. Where applicable, terms and names in the database are also geo-referenced.

Keywords Thesaurus; cultural heritage resources; content management platform; controlled vocabularies; bilingual resources

1. Introduction

Cultural tourism tends to be a learning-intensive experience, in the sense that visitors most often wish to

> learn, discover, experience, and consume the tangible and intangible cultural attractions and products in a tourism destination. These attractions/products relate to a set of distinctive material, intellectual, spiritual, and emotional features of a society that encompasses arts and architecture, historical and cultural heritage, culinary heritage, literature, music, creative industries and the living cultures with their lifestyles, value systems, beliefs, and traditions.[1]

In this regard, the project seeks to address the visitors' needs to learn and experience virtual itineraries in Eastern Macedonia and Thrace, Greece, by developing a platform that integrates cultural heritage content relative to this area. In the paper, we present work aimed at building a thesaurus that is tailored to meet the needs of indexing and retrieval of the cultural heritage content in the resulting platform.

The paper is structured as follows: In section 2 we will give a brief account of the project in terms of its scope, aims, and expected results. Then, the cultural content will be presented in section 3 along with the web interface that serves as a content management system. The focus will be on the description of the thesaurus developed, the design principles it abides to, as well as the twofold purpose it serves in the project and the final platform (section 4). We will then elaborate on the methodological approach taken for creating the thesaurus and

[1] This definition was adopted by the United Nations World Tourism Organization General Assembly, at its 22nd session (2017).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

controlled vocabularies (section 5). In section 6 we will present background on thesauri and controlled vocabularies aimed at cultural heritage content management. Finally, our conclusions and plans for future research will be outlined in section 7.

2. Project aims and scope

Mythotopia (short for "Mythological routes in East Macedonia and Thrace") is currently a Research & Development project in progress, aimed at developing an online platform, which offers a multi-faceted view of Eastern Macedonia and Thrace, Northern Greece. This includes a wealth of information from several points of view including mythology, history, architecture, natural environment, culture, society, folklore, recreation, gastronomy, travel and tourism, and leisure. The primary aim of Mythotopia is to bring to light the cultural wealth of the region in the form of a bilingual Greek (EL) and English (EN) informative platform, in view of eventually contributing to its tourism development (Vacalopoulou et al. 2021). In order to achieve this aim, Mythotopia records, maps, and highlights local cultural and tourist attractions under the scope of mythology. Users of Mythotopia will be given the chance to combine elements of interest connected to the current natural, social, and cultural landscape with the rich mythological background of the area to create their preferred virtual routes. They can, then, use their selected routes and the extensively researched accompanying information as guides in actual visits on site. By combining these elements, Mythotopia attempts to join the past and the present to help create a complete tourist experience (ibid.).

3. The cultural tourism ecosystem: the content and web application

Project development is centred around two main pillars: (a) the cultural content relative to the areas of Eastern Macedonia and Thrace, which is multi-faceted and highly heterogenous in nature, and (b) the resulting platform, that is, a Graphical User Interface (GUI) for integrating, documenting, geo-tagging, storing, and retrieving data based on specific metadata elements. In between, the thesaurus serves as the backbone of the platform and helps indexing and retrieving the content.

3.1 The platform content

Following project specifications, the data types stored and, therefore, documented in the platform comprise: (a) myths that are relevant to the area of Eastern Macedonia and Thrace; (b) Points of Interest (POIs), that is, a variety of tangible and intangible elements that are characteristic of the area pinpointing its cultural, social, and environmental identity; and (c) audio-visual material.

To start with, ancient Greek myths related to the area are the primary data to be stored and documented in the platform in the form of narratives. These myths are centred around specific living entities, as for example deities, heroes, and mythological creatures; moreover, they are linked to specific localities (places), religious or ritual practices, and the culture of the time. Ancient myths have been made known via the seminal literary works of ancient Greek and Latin authors, and they have been also referred to in travel memoirs of the past.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Therefore, a significant part of the mythological component consists of ancient Greek and Latin literary texts featuring myths or mythological figures and localities that are linked to the area. The literary texts are excerpts that were selected from established scholarly editions; they cover a variety of genres and a wide range of Ancient Greek and Latin literary production both in prose (historiography, myth-writing, biography, rhetoric, philosophy, and scientific texts, such as geographical works and ancient scholia) and in verse (epic, drama, elegy, epigram, lyric poetry, bucolic poetry, and didactic poetry). Moreover, texts that pertain to the genre of travel literature, that is, travel writings produced by authors who visited the area of Eastern Macedonia and Thrace both in ancient times (i. e., Pausanias, Strabo), and in Medieval and Modern times (i. e., Buondelmonti, Ciriaco d'Ancona) were also selected for inclusion. The Ancient Greek and Latin literary texts are stored in the original; consequently, their translations in EL and EN are also provided by the project participants. Additionally, accompanying textual material in the form of biographies of authors, general information about the mythological figures is also developed by the project participants and included in the raw data. Finally, the corpus comprises narrative texts in EL and their translations in EN depicting the myths that are of relevance to the broader area of interest. To date, a total of c. 300 texts has been collected or produced.

Along the same lines, visitors may gain valuable insights of the culture and myths related to the area by means of artifacts of all sorts. Thus, images of clay pots, sculptures, engravings, coins, various metal and glass objects, mosaics, sarcophagi, etc. portraying figures, or scenes of the myths, dating back to Greek and Roman antiquity, were also selected. In addition, the collection includes references to excerpts of classical operas depicting mythological scenes and stories. Finally, informative texts, accompanying images, and videos tailored to meet the needs of tourists visiting the area have also been included in the collection. In terms of content or subject matter, the textual and audio-visual content features primarily entities of the following types: living entities (i. e., animals, plants that are endemic in the area), geopolitical entities (i. e., cities, towns, villages, or minor settlements), geographical entities (e. g., mountains, rivers, beaches, lakes), facilities and archaeological sites, events, and activities, as well as intangible cultural elements (i. e., food and gastronomy, folklore, and cultural events of the area). In essence, these entities constitute the body of POIs of the resulting application.

Besides the various annotations applied to the data (Giouli et al. 2022), the development of a taxonomy of object types was of paramount importance in view of better accounting for the heterogeneous dataset stored in the project database. At the same time, according to the specifications of the project, the informative material that was initially prepared in Greek was to be translated to English. The controlled vocabulary that was developed also helped in the translation process, as will be shown in section 4.1 below. Therefore, the purpose of thesaurus creation was two-fold: on the one hand, it was aimed at documenting and indexing the primary data, facilitating retrieval over the database constructed; and on the other hand, it assisted translators towards ensuring consistency throughout the translation phase.

3.2 Collaboratively storing and managing data: the database and web application

A web platform was developed as the front-end of a database for collaboratively storing, documenting, and searching the cultural content. The database supports various modalities (text, images, audio, and video) in two languages: EL and EN. In addition, it was designed to

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

meet the following requirements as a minimum: (a) user friendliness for both people who are responsible for data entry and the end users who will use the taxonomy for browsing and retrieving information; (b) extensibility, that is, the taxonomy should be open and easily modified or extended in view of incorporating new datatypes in the future; (c) functionality, that is, the thesaurus and the overall platform should provide certain functionalities for inter-linking entries, as well as managing the workflow from the initial entry of data to the final publishing of entries; and (d) scalability, the system must deal with (relatively) large amounts of data, still offering a friendly environment. Therefore, the platform was designed so as to meet a series of crucial requirements that support end outcome quality and user satisfaction.

Items stored in the platform (myths, artefacts, POIs) were indexed with respect to a location term provided in the thesaurus. In addition, one novel feature of the digital collection lays in the fact that part of the cultural content has also been geotagged manually. More precisely, POIs have been assigned geospatial metadata in the form of geographical coordinates using OpenStreetMap and Leaflet,[2] a light, open-source JavaScript library that works across all major desktop and mobile platforms.

The web application is designed to provide access to the main taxonomy objects in an order of importance as a part of a workflow from the main taxonomy to secondary ones. As a starting point, myths constitute the main or central taxonomy; therefore, they are managed first. All the other taxonomies (literature, media, POIs) are imported and connected to the main taxonomy. However, during the data entry period, a more myth-independent approach was also employed, and secondary – or less central - taxonomies were submitted to the database on their own, and then editors provided inter-connections. This proved to be faster and more efficient in creating and managing the final collections.

4. Thesaurus description

In this section, we will elaborate further on the thesaurus that supports this application. The major issue the project team had to cope with when defining the taxonomy, was the heterogenous nature of the dataset at hand. To overcome this challenge, we initially identified the subject fields or domains that different types of data fall into; each domain was treated as a separate thesaurus thereof with inter-dependencies where applicable. Thus, the following fields were identified: literature, archaeology, arts, and travel. The backbone of the thesaurus is an upper-level taxonomy that starts from the aforementioned fields and depicts a set of classes and subclasses that correspond to the concepts of the taxonomic schema. Ultimately, the concepts or classes are being populated with domain-specific terms.

Following standard specifications for thesauri creation (Aitchison/Gilchrist/Bawden 2000), a set of relations have also been defined that hold between concepts, between concepts and terms or between terms. Thus, three types of relations are foreseen: semantic, spatial, and associative ones. These ultimately create a semantic network that shows links and paths between terms and concepts. The standard semantic relations that have been defined, namely the hierarchical relation (between concepts) and the equivalence one (between terms), designate the vertical and the horizontal axes of the taxonomic schema respectively. The former deals with defining the hierarchical structure of the thesaurus. Concepts at the upmost level define the basic categories or classes of cultural data objects. A total of 37 classes have been

[2] https://leafletjs.com/.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

defined in our schema; these are instantiated as the Broad Terms (BTs) of the taxonomic schema. The hierarchical relation in the thesaurus links these concepts with their subordinate ones, the so-called Narrow Terms (NTs), establishing thus the hierarchical structure in a thesaurus. In total, 144 subclasses have been stipulated, which are further grounded to the terms of the thesaurus.

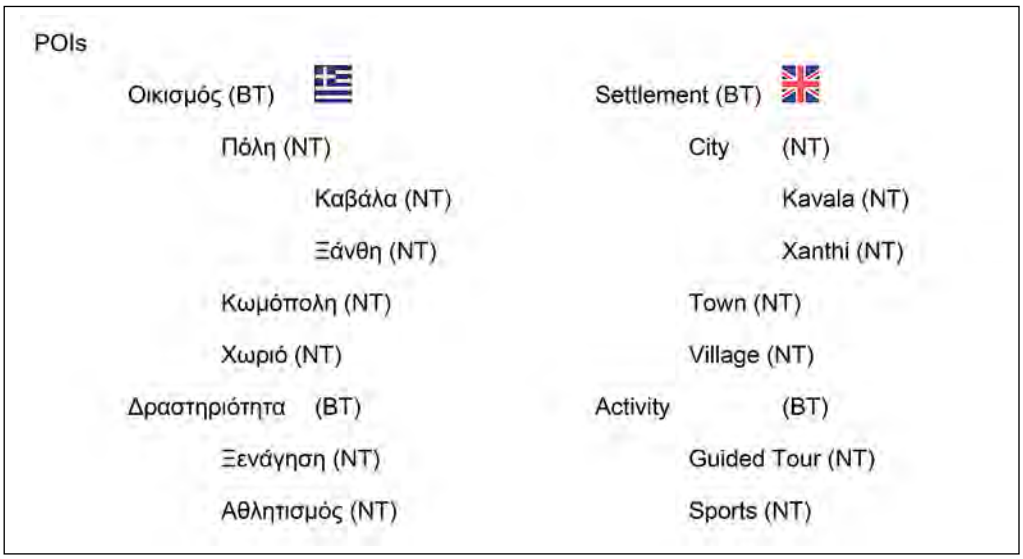


Fig. 1: Snippet of the hierarchical structure of the thesaurus

Concepts at the first two upmost levels are unambiguously grounded to terms in Greek and English. Moreover, guidelines elaborated in the framework of the project further specify the content of these concepts. For example, types of settlements are distinguished with respect to the number of inhabitants, whereas official records are provided to the people involved in the task of data entry and indexing. This is not the case with the concepts at the downmost level of the hierarchy. For these concepts, a preferred term is employed assuming the role of the descriptor. According to the specifications set, the descriptor is the unmarked form of a term, that is, one that pertains to general language, as opposed to dialectical or otherwise marked terms. Historical and geographical variations of terms are also encoded in the thesaurus as alternative terms or non-descriptors. The equivalence relation is then established between a descriptor and one or more non-descriptors. This relation is of particular importance for the data at hand and the final application since it is used in computing itineraries. The thesaurus encompasses more than 1820 terms, of which c. 1,400 are descriptors. Apart from terms, a set of 139 keywords or key phrases featuring myth-related concepts have been so far included in the thesaurus for better indexing the mythological narratives.

In addition, associative relations are also specified in the thesaurus. The generic relation RelatedTo (RT) is further specialised in a set of sub-relations, each one specifying the items related. Therefore, the relation RelatedTo-POI is used to link an item documented in the platform with one or more POIs in the database; the relation RelatedTo-image (RT-image) is used to link an item with an image, whereas RelatedTo-file and RelatedTo-author are used to link myths with literary and travel texts and corresponding authors respectively. Finally, spatial relations are also foreseen. More precisely, the relation Area links one of the regional

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

units of the area with the POIs in the database. Where applicable, terms in the database are also geo-referenced. Figure 2 illustrates a snippet of the model for the description of POIs.

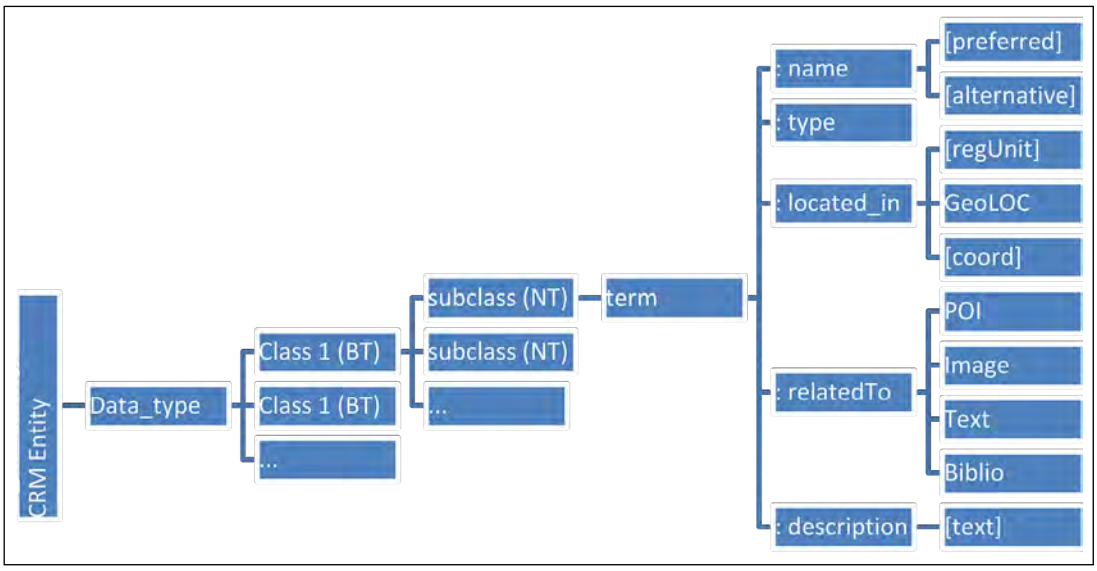


Fig. 2: A sample of the taxonomic schema with relations

We opted for establishing hierarchical and associative relationships between concepts rather than between terms, since this has been proved as a prerequisite for thesaurus interoperability (Dextre Clarke/Zeng 2012).

Apart from common nouns relating to concepts and terms, the thesaurus also includes proper nouns that is, names of mythological characters of the area of Eastern Macedonia and Thrace (heroes, gods, etc.), as well as names of artists, and place names (i. e., names of geopolitical and geographical entities). Historical and dialectal variants of toponyms are also included as terms; moreover, dialectal as well as scientific names of entities (i. e., animals and plants) are also retained.

4.1 Using controlled vocabularies to bridge languages and domains

To achieve the highest possible level of control, standardisation and homogeneity throughout the collection, detailed guidelines were elaborated specifying linguistic choices and presentation. These include choice and ordering of terms, selection of descriptors over non-descriptors, orthographic conventions, grammatical forms, capitalisation policy, abbreviations, and acronyms, as well as usage of punctuation marks.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

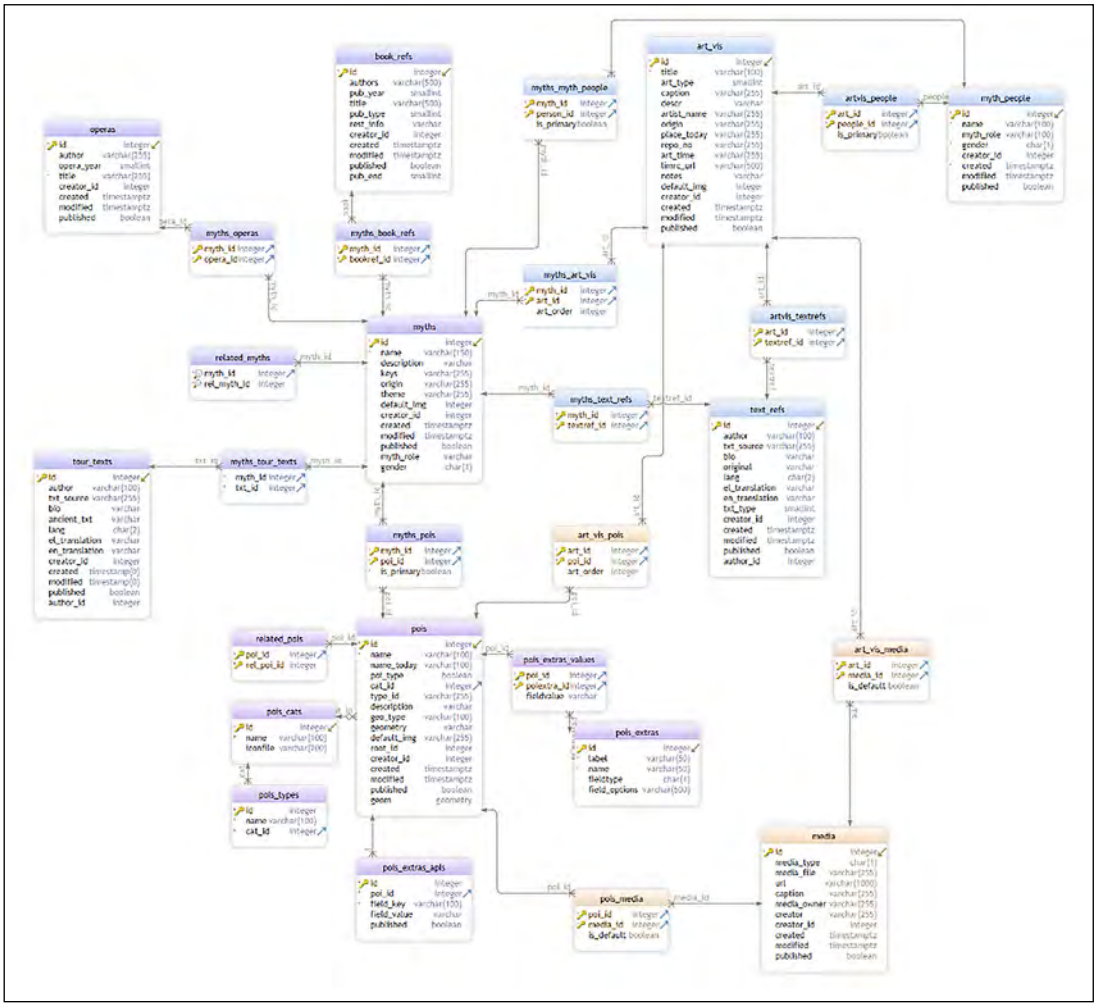


Fig. 3: Interlinking of entities springing from the central "myths" table in a small snapshot of the thesaurus architecture

Interlinking of data or entries in the platform via the relations of terms in the thesaurus results in bridging domains which seemed unrelated. For example, the myth of Voreas in the mythological component of the platform is linked to literary works (Literature) and images of Archaeological artifacts (Archaeology). By linking it to the item Pangeon (a mountain that is documented in the Travel domain), a bridge is established between the two domains (Travel and Literature). This is depicted in Fig. 3, in which the myth of Voreas belongs in the central "myths" table, which is linked through intermediate tables to "text_refs" (where the respective literary works are stored), to "art_vis" (where images of related artifacts are kept), to "pois" (populated by related POIs, in this case, Pangeon, a geographical element and the respective bridge, a transport element).

From another perspective, the controlled vocabularies were also consulted when translating the platform content from Greek to English in an attempt to achieve accuracy as well as consistency throughout the translated content. As a first step, the establishment of translational equivalents between the controlled vocabularies in the two languages was in order. This has not always been an easy task, especially in the case of translating specialized terms, as for example those referring to plant and animal species and sub-species. In this case, the translation of the descriptors was based on non-descriptors, namely, the scientific names in Latin.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

5. Methodology for thesaurus construction

The main activities within the project life cycle can be outlined as follows: 1) identification and road mapping of the data types that will be stored in the platform (myths, POIs, audio-visual material) and the sources for acquiring them; 2) definition of the metadata elements that are appropriate for documenting each one of them; 3) selection of the appropriate concepts and terms that were relevant to the dataset, and 4) building the platform and deciding upon its functionalities taking into consideration not only the people who are responsible for data entry, but also the prospective users of the platform.

The project team opted for a hybrid approach to thesaurus creation: a top-down development phase was complemented by a bottom-up one. In our top-down approach, a list of seed terms that roughly correspond to our taxonomy was initially postulated. Using the textual data that was collected and produced within the project, we then populated the classes with instances; revisions and extensions of the initial schema were made where needed. For our bottom-up procedure to thesaurus creation, the textual data collected served as a corpus; terms were selected for inclusion from the data. The terms were mined from reliable sources that had been selected early in the project life cycle. Our work then involved defining the hierarchy and rest of relations, designating the preferred term or descriptors of the concepts and its variant term(s). Following standard methodologies, and previous best practices for thesaurus creation (Aitchison/Gilchrist/Bawden 2000), a set of guidelines were implemented to ensure the usage of controlled vocabularies throughout the lexicographic process.

The thesaurus broadly covers the domains of Archaeology, Literature, Mythology, and Travel. From another perspective, it can be viewed as a set of inter-linked thesauri. The main challenge that we had to deal with when constructing the thesaurus, however, was the high degree of fragmentation between the various domains and datatypes that we needed to include in the platform. In this respect, we tried to be inclusive and develop self-sustained thesauri that would be adequate for each domain, allowing for overlapping concepts/classes across domains if needed.

6. Background on thesauri and controlled vocabularies

Over the last decades, thesauri have been developed and utilized by digital content holders and providers, such as digital libraries, archives, and museums. These are mainly characterized by a considerable degree of specialization and granularity in distinguishing and layering concepts. For example, UNESCO Thesaurus[3] is a controlled and structured list of terms used indexing and retrieval of publications in seven subject fields or domains, namely education, culture, natural sciences, social and human sciences, communication, and information; each subject field is broken down further into micro-thesauri which allow users to gain a quick overview of the subject matter. With its first release dating back in 1977 in English, the thesaurus has over the years been extended to also include French, Spanish and Russian. Similarly, the CIDOC Conceptual Reference Model (CIDOC CRM)[4] is a theoretical and practical tool for information integration in the field of cultural heritage (Bekiari et al. 2021). The Getty Vocabularies contain structured terminology for art, architecture, decorative arts, archival materials, visual surrogates, art conservation, and bibliographic materials. Compliant

[3] https://skos.um.es/unescothes/.

[4] https://www.cidoc-crm.org/Version/version-7.2.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

with international standards, they provide authoritative information for cataloguers, researchers, and data providers. However, the decision for building our own taxonomy instead of adopting an existing one was essentially directed by the nature of our data. Being highly heterogenous, not only in form (texts, video, audio) but also in terms of the domain (or even the genre in the case of texts) they pertain to, they were difficult to organize.

7. Conclusions and future research

We have presented a thesaurus that has been elaborated in view of indexing and retrieval of cultural heritage content over a dedicated platform aimed at enhancing tourists' experiences in the area of Eastern Macedonia and Thrace. The core of the thesaurus is a collection of concepts represented by terms and interlinked by relationships. These concepts pertain to the domains of Archaeology, Literature, and Travel. Terms in the thesaurus have already been employed to index text, video and audio that is featured on the touristic platform. Future work has been already planned towards enriching the thesaurus with new terms. Moreover, to render our taxonomic schema compatible with existing digital libraries and other state-of-the-art conceptual models (i. e., CIDOC CRM), mappings from our schema to standardized ones are underway. The resulting platform and the thesaurus will be freely available and accessed over the web.

References

Aitchison, J./Gilchrist, A./Bawden, D. (2000): Thesaurus construction and use: a practical manual. 4th edition. London.

Bekiari, C./Bruseker, G./Doerr, M./Ore, C.-E./Stead, S./Velios, A. (2021): Definition of the CIDOC Conceptual Reference Model. Version 7.2. https://www.cidoc-crm.org/sites/default/files/cidoc_crm_version_7.2.pdf (last access_ 20-06-2022).

Dextre Clarke, S. G./Zeng, Marcia Lei (2012): From ISO 2788 to ISO 25964: the evolution of thesaurus standards towards interoperability and data modelling. Information Standards Quarterly 24, 1 (Winter), pp. 20–26.

Garshol, L. M. (2004): Metadata? Thesauri? Taxonomies? Topic maps! Making sense of it all. In: Journal of Information Science 30 (4), pp. 378–391.

Giouli, V./Vacalopoulou, A./Sidiropoulos, N./Flouda, C./Doupas, A./Giannopoulos, G./Bikakis, N./Kaffes, V./Stainhaouer, G. (2022): Placing multi-modal, and multi-lingual data in the humanities domain on the map: the Mythotopia Geo-tagged Corpus. In: Proceedings of the 13th Chapter of the International Language Resources and Evaluation Conference (LREC 2022), 20–25 June 2022, Marseille, France.

Nielsen, M. L. (2004): Thesaurus construction: key issues and selected readings. In Cataloging & Classification Quarterly 37 (3–4), pp. 57–74.

Ryan, C. (2014): Thesaurus construction guidelines: an introduction to thesauri and guidelines on their construction. Dublin. DOI: 10.3318/DRI.2014.1.

UNESCO and International Bureau of Education (IBE) (1984): IBE education thesaurus: a list of terms for indexing and retrieving documents and data in the field of education – with French and Spanish equivalents. Paris.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Vacalopoulou, A./Mastrogianni, A./Michalopoulos, C./Tsiafaki, D./Michailidou, N./Mourthos, I./Botini, P./Stainhaouer, G. (2021): Mythological itineraries along the western silk road: finding myths in visits to eastern Macedonia and Thrace today. In: Silk road sustainable tourism development and cultural heritage. The University of Thessaloniki and European Interdisciplinary Silk Road Tourism Centre.

Contact information

Voula Giouli
Institute for Language and Speech Processing, ATHENA RC
voula@athenarc.gr

Anna Vacalopoulou
Institute for Language and Speech Processing, ATHENA RC
avacalop@athenarc.gr

Nikos Sidiropoulos
Institute for Language and Speech Processing, ATHENA RC
nsidir@athenarc.gr

Christina Flouda
Institute for Language and Speech Processing, ATHENA RC
cflouda@athenarc.gr

Athanasios Doupas
Institute for Language and Speech Processing, ATHENA RC
adoupas@athenarc.gr

Gregory Stainhaouer
Institute for Language and Speech Processing, ATHENA RC
stein@athenarc.gr

Acknowledgements

We acknowledge support of this work by the project "Mythological Routes in Eastern Macedonia and Thrace" (MIS 5047101), which is implemented under the Action "Regional Excellence", funded by the Operational Programme "Competitiveness, Entrepreneurship and Innovation" (NSRF 2014-2020) and co-financed by Greece and the European Union (European Regional Development Fund).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Anke Müller/Gabriele Langer/Felicitas Otte/Sabrina Wähl

CREATING A DICTIONARY OF A SIGNED MINORITY LANGUAGE

A bilingualized monolingual dictionary of German Sign Language

Abstract Lexicographers working with minority languages face many challenges. When the language in question is also a sign language, circumstances specific to the visual-spatial modality have to be taken into consideration as well. In this paper, we aim to show and discuss which challenges we encounter while compiling the Digitales Wörterbuch der Deutschen Gebärdensprache (DW-DGS), the first corpus-based dictionary of German Sign Language (DGS). Some parallel the challenges minority language lexicographers of spoken languages encounter, e. g. few resources, no written tradition, and having to create one dictionary for all potential user groups, while others are specific to sign languages, e. g. representation of visual-spatial language and creating access structures for the dictionary.

Abstract Sign language dictionary; minority language; bilingualized dictionary

1. Introduction

This contribution discusses issues concerning the creation of dictionaries for signing minorities with a focus on the German Sign Language (DGS) community and shows possible solutions to the challenges at hand. Sign languages in general are lesser-resourced languages. They are embedded in hearing-majority communities, and deaf communities have developed their own visual languages and cultures. Only recently has the collection of sign language (SL) corpora become technically feasible, because digitalization has made it possible to collect, store and process big amounts of video data. The DGS corpus was compiled for lexicographic and documentation purposes. It is the foundation for the *Digitales Wörterbuch der Deutschen Gebärdensprache – Das korpusbasierte Wörterbuch DGS* (*DW-DGS*). This will be the first dictionary of DGS that is corpus based and probably will stay the only one for a while. Like other minority language dictionaries, the DW-DGS will need to serve different user groups at once.

This text contains a theoretical part giving background information on minority languages and SLs in particular, as well as a brief overview on SL lexicography. In the second part, challenges and the respective solutions for the DW-DGS are discussed.

2. Background

2.1 Sign languages

Sign languages are visual-spatial languages and use several channels (hands, face, mouth, body posture, eye-gaze) at the same time for communication. With multiple articulators available, SLs allow for a high degree of simultaneity. The lexical meaning usually is carried by the manual component of signing sometimes in combination with a mouthed word (mouthing) whereas non-manuals often but not exclusively carry grammatical information.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Manual signs together with other bodily actions can be used to create productive and often iconic signs. Accordingly, SLs allow for a high degree of iconicity. Due to these characteristics, SLs differ greatly from the surrounding spoken majority languages.

SLs exhibit a high level of variation and fast-paced language change. While politically or educationally motivated discussions concerning standardization of SLs exist on the one hand, on the other hand the language community itself highly values the regional variation of signs. To date, no standard for DGS has evolved, though levelling effects may increase in the future due to higher mobility and facilitated inter-regional exchange by technical means like video chat.

SL research is a relatively new field in linguistics, dating back into the 1960s for the USA and the 1980s for Germany. In the beginning stages, scientists were mostly concerned with proving that SLs are natural languages as they were for a long time disregarded as primitive gestural and non-linguistic systems (cf. McBurney 2006, p. 311). Since then, SL research has come a long way and a considerable number of SLs have been recognized as languages by the respective governments. New technologies, especially digitalization and the possibility to record and store large amounts of video data, strengthened SL research.

Due to the lack of a widely used writing system, film is still the preferred means for capturing, documenting, and representing SLs. Technologies based on writing cannot be applied to SLs, and annotation usually has to be done manually. SL dictionaries thus need to find a way to represent and describe the SL adequately (more on this in chapter 3.3).

2.2 Minority languages – spoken and signed

The idea of majority and minority languages is closely tied to the development of nation states, where often only one language was selected as national language (Wright 2018). People not speaking that national language became minorities, a process which entailed negative attitudes towards the minority language and little institutional support in many cases (Dorian 2006, p. 440). First international laws to ensure non-discrimination of minority members were established starting in the late 1950s, but only in 1992 would the UN declaration on the Rights of Persons Belonging to Ethnic, Religious and Linguistic Minorities contain requirements for governments "to promote as well as protect the identity of minority groups" (Wright 2018, p. 644).

In comparison to majority-language communities, linguistic minorities have a smaller number of language users. Minority languages are also usually spoken in a specific area of a country. Most indigenous languages that are considered minority languages do not have a written tradition, and transmission is provided by face-to-face interactions from one generation to another.

Deaf communities communicating in signed languages exist all over the world, and generally constitute linguistic minorities within their countries. These minorities do not differ from the population by territorial or ethnic origin but evolved from deaf people gathering and finding their own way of communication – a visual language with a visual culture. This view of the deaf community as a linguistic minority and not a group connected via disability and hearing loss has arisen in the 20th century, fought for by deaf activists and allies. In the wake of these struggles, legal language recognition of SLs started in the 1990s (De Meulder/Murray/McKee 2019).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Sign language minorities differ from other linguistic minorities by the way of language transmission and the constitution of its members. There is only minimal parent-to-child transmission, as over 95 percent of deaf children are born to hearing parents and thus do not have immediate primary language access to a signed language (Hartman/Nicolarakis/Wang 2019, p. 3; Plaza-Pust 2016, p. 18). Generally, they acquire a SL from their peers at school or once they become acquainted with signing people (Plaza-Pust 2016). The spoken majority language is not fully sensorily accessible to deaf or hard-of-hearing children, even if they are equipped with hearing assistance systems such as digital hearing aids or cochlear implants from early on (Hartman/Nicolarakis/Wang 2019). This demographic situation may account for a comparably late SL acquisition of many deaf children, sometimes following an unsuccessful trial of spoken language acquisition (ibid.). In the context of SLs, the term L1 denotes the SL as the language of choice that is fully visually accessible and satisfies expressive needs. It is associated with a sense of belonging and emotional connectedness. It is the language of in-group everyday use rather than the language that was acquired first.

2.3 DGS language community

Sign language communities are heterogeneously composed. They comprise all people who know and use the local SL and belong to the visual culture; this includes hearing children of deaf parents and hearing partners who sign, as well as deafened persons who sign. To be deaf or hard of hearing is neither a prerequisite nor sufficient to belong, and there are deaf/hard-of-hearing people who have not acquired SL or who prefer oral and written communication in the majority language.

In Germany, DGS was recognized in 2002 within a law concerning disability rights.[1] It does not have the status of an official language, but its status provides the legal framework for some accessibility measures (e. g., signed video translations of governmental internet sites). The recognition certainly has improved the situation of deaf persons, but the consequences of decades of language suppression especially through the educational system are still present (Plaza-Pust 2016). It was and is believed by some scholars that learning a signed language would prevent or impede the acquisition of spoken language (cf. Hall/Hall/Caselli 2019). The "oral method", focusing on spoken language only, was installed at boarding schools, and bilingual schooling in German and DGS only started in the early 1990s in experimental settings (Plaza-Pust 2016). Today, bilingual education "continues to represent the exception rather than the norm" (ibid., p. 449). Another factor impeding the transmission of SL today is the closure of special schools for the deaf in favor of inclusion or mainstreaming, leading to the isolation of deaf individuals and, in consequence, language deprivation.

Members of the signing community live in a permanent language contact situation, where speaking and writing in the majority language prevails. German is taught in school and is the language of education. Consequently, German signers are bilingual to a certain degree. Print literacy in the majority language is not easily acquired for deaf people. Nevertheless, deaf people use writing of the majority language in text-based communication devices, e. g. through SMS or e-mail for communication with hearing persons as well as among themselves (Power/Power 2004; Maxwell 1985) and as a means of taking notes. Also, deaf people

[1] See BGG, §6 https://www.gesetze-im-internet.de/bgg/index.html#BJNR146800002BJNE000601119; §6 https://www.gesetze-im-internet.de/bgg/__6.html.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

are used to subtitled television ("caption literacy" according to Paul 2018) and thus are acquainted with the semantics of German and the written language of the majority. Consequently, the deaf community encompasses a wide range of reading and writing skill levels.

Studies on literacy of deaf people have found that reading comprehension and writing skills are often delayed or below average, concluding that many deaf people are functionally illiterate (cf. Harris/Terlektsi 2021, p. 12). Adopting a different, skill-oriented perspective and taking into regard the daily means of communication (print, electronic media, subtitled videos) along with signing, we would rather talk of a functional bilingualism observable in deaf individuals, instead of highlighting underachievement.

2.4 Sign Language lexicography

Lexicography of under-researched and lesser-resourced languages, especially those without a written tradition, has to cope with a number of circumstances: written sources are scarce or missing altogether, corpora are small or non-existent, language-specific NLP such as automatic PoS taggers and lexicographic word profile tools are often not available. Exclusive face-to-face language use and previously undocumented language structures pose challenges to researching and describing sentence grammar including boundaries, syntax, and PoS-categories. Often lexicographers cannot build on pre-existing comprehensive and commonly accepted descriptions of their object language's grammar. Generally, all of these challenges also apply to SL lexicography.

In addition, SL lexicography has to deal with the circumstance that it cannot resort to an established and adequate writing system. This makes corpus design, annotation, as well as corpus analyses for lexicographic descriptions difficult and cumbersome. Only recently have corpora become available. As a consequence, working corpus-based and applying practices and methods from spoken language lexicography, e. g. using concordances, collocational analyses, and the like for word sense discrimination is rather new to the field and only just evolving (see for example Langer/Müller/Wähl 2018; Langer/Schulder 2020).

For a dictionary design not having a writing system means in essence that a written representation of signs is not available to represent example sentences, and to function as the guiding and ordering elements (lemmata, elements representing cross-reference addresses) in the macro- and microstructure of the dictionary. Related to this question is the issue of ordering or – in electronic dictionaries – searching via sign form in order to enable a bidirectional use also from sign to word in bilingual dictionaries.

Modern SL lexicography is a field that is still new and developing. It is dedicated to finding adequate solutions for the challenges SLs present as object languages. The *Dictionary of American Sign Language on Linguistic Principles* (DASL) (Stokoe/Casterline/Croneberg 1965) marks the beginning of modern SL lexicography. The DASL was the first profound SL dictionary that took the manual signs of a SL as basis for their lexicographic description of their properties and looked at units of ASL from a monolingual perspective. It provided a search by phonological parameters (such as movement, handshape, hand orientation, and location) through the macrostructure via notation and included information on meanings and usage. Stokoe/Casterline/Croneberg were also the first to attempt to base their lexicographic descriptions on a corpus of filmed signing.

SL dictionaries prior to the DASL – but also many SL dictionaries of more recent times – are bilingual unidirectional sign collections pragmatically compiled on the grounds of intro-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

spection. They basically consist of spoken language word lists combined with matching form representations of sign equivalents. Usually, no further information on the signs' other properties is provided. Collections of this kind tend to present one sign equivalent for each word – thus implying a 1:1 relationship of words and signs and very rarely do they offer searchability for sign form (Stokoe 1993, p. 138; Zwitserlood/Kristoffersen/Troelsgård 2013, pp. 260 f.).

SL dictionaries following the approach layed out by the DASL on the other hand look very different as they focus on the signs and their meanings and properties from a monolingual perspective. Several of this kind have been produced since – e. g. Johnston (1998) for Auslan, ODT for Danish SL. In recent larger dictionary projects such as the CDPSL and the DW-DGS, new possibilities are being explored of how to base entry information on corpus data. As the electronic medium invites the inclusion of signed information recorded on film, new dictionary structures are being developed to integrate these information types alongside with written information into entries and access structures.

2.5 User groups of minority language dictionaries

As funds are scarce, dictionary makers for minority languages, who tend to be non-native speaker linguists, missionaries, or members of the language community (Bradley 2015), get one shot to complete a one-size-fits-all dictionary without the prospect of being able to update it (Cristinoi/Nemo 2013). They have to serve as many user groups as possible despite their diverse needs: the language community, language learners (Prinsloo 2012), and the academic community (Mosel 2004). In the case of signed minority languages, this list can be extended to include language professionals such as SL teachers and interpreters, hearing people in contact with deaf signers including those with close contact such as hearing parents of deaf children, service providers to the deaf community, and finally the interested public (compiled from Moskovitz 1994; Hilzensauer 2000; Vale 2015; McKee 2017). Such a dictionary has many purposes (Cristinoi/Nemo 2013): research, documentation, preserving linguistic and cultural heritage, helping native speakers communicate in the dominant language, helping learners, and, at least for spoken languages, providing a stable orthography.

Different user groups have diverse needs and wishes regarding the features of a SL dictionary. Features frequently discussed in this context include the ordering of entries, the mode of sign depiction (picture, drawing, video), the type(s) of search function, and the different information types included in an entry (Hilzensauer 2000; Moskovitz 1994). Unfortunately, published surveys or studies on SL dictionary usage are few and far between. Table 1 summarizes some of the results of three studies. Moskovitz (1994) is a questionnaire inquiring about participants' preferences and expected usage ahead of the construction of a dictionary. Kristoffersen/Troelsgård (2012a) is a questionnaire for users of a published online dictionary of Danish Sign Language (DTS). Vale (2015) is a study of the actual usage of an online dictionary using log files and interview data.

The dictionary is used the most for…	…learning SL, as a quick reference (Moskovitz 1994) …learning SL, for fun (T. Troelsgård, personal communication, March 2022) …preparing for specific communicative situations (Vale 2015)

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

The dictionary is most used in the direction of…	…English to NZSL (Moskovitz 1994) …Danish to DTS (Kristoffersen/Troelsgård 2012a) …English to NZSL (Vale 2015)
The preferred information types are…	…sign grammar, synonyms, production instructions, English usage examples (Moskovitz 1994) …sign videos, signed example sentences (Kristoffersen/Troelsgård 2012a) …sign videos (Vale 2015)
The lesser used/desired information types are…	…English phonetics, information on other SLs (Moskovitz 1994) …hyperlinks: synonyms, concordances, information on the Danish words (Kristoffersen/Troelsgård 2012a) …hyperlinks (Vale 2015)

Table 1: Summary of three surveys/studies on SL dictionary usage

Based on these results, SL learners are SL dictionaries' first and foremost user group. Searches are usually carried in the direction from spoken language to signed language. However, both of these findings may change over time, as deaf people become more experienced in the handling of dictionaries and all potential users learn how to use sign-based search functions, which can seem unusual and intimidating at first. Regarding the entries' contents, the focus is on information specific to the given sign rather than on other languages such as the surrounding spoken language. Interestingly, Kristoffersen/Troelsgård (2012a) point out that the lesser used information types in the ODT are presented in the form of links. Vale (2015) also finds that hyperlinks in entries are seldomly clicked on. This shows that the form of presentation of the information is very influential and that information that is not directly embedded in the entry is less likely to be looked at by the user.

3. DW-DGS

The *Digitales Wörterbuch der Deutschen Gebärdensprache – Das korpusbasierte Wörterbuch DGS – Deutsch* (*DW-DGS*) is currently being compiled based on the data of the DGS corpus. Dictionary and corpus are produced by the DGS-Korpus project (http://dgs-korpus.de). The DGS corpus is the largest corpus available for DGS and has reached a substantial size of more than 670.000 tokens (as of 2022-03-25). It was designed to serve as a general reference corpus for DGS and to provide data for the compilation of the DW-DGS.

3.1 Dictionary type

The DW-DGS is a corpus-based, monolingually oriented bilingualized general descriptive dictionary of DGS. For the first time, DGS signs, their meanings, grammatical properties, and usage patterns can be studied in their linguistic context with a corpus linguistics approach. Working corpus-based is a requirement for reliable lexicographic descriptions (cf. Atkins/Rundell 2008, pp. 53 f.). Thanks to the robust evidence that the corpus provides for variation in DGS, the DW-DGS is able to include both lexical and phonological variation and thus avoids being a standardizing influence as much as possible.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

The DW-DGS takes a monolingual perspective in that it focuses on DGS, the minority language, as the object language of interest. Creating a monolingual dictionary for an under-resourced language is a means to support self-acknowledgement, empowering the members of a cultural and linguistic minority, and promoting a positive attitude towards the minority language within the majority group (Erlenkamp 1998 for DGS). To base lexicographic descriptions on corpus data collected from members of the signing community is a prerequisite for adequateness and reliability of dictionary contents and, through participation, respects and reflects the cultural status of DGS within its minority language community. To raise accessibility and impact, the DW-DGS is not only a monolingually focused dictionary but includes some bilingual features. It provides German translation equivalents and a German index.

Like most other general SL dictionaries, the DW-DGS has opted for a hybrid version between a monolingual and bilingual dictionary. It could be called a *bilingualized monolingual dictionary* or, for having the definitions given in German, a *bridge dictionary* or a *semi-bilingual dictionary*. Such a hybrid approach is not unusual for SL dictionaries, see for example the online Danish Sign Language Dictionary (cf. Kristoffersen/Troelsgård 2012b, p. 302). Such a bilingualized dictionary is, generally speaking,

> one that offers T[arget]L[anguage] equivalents while retaining the S[ource]L[anguage] definitions from the monolingual dictionary on which it has been founded. Such dictionaries are always monodirectional and monoscopal (L2-L1), with only an L1-L2 index in place of a regular L1-L2 section. (Adamska-Sałaciak 2013, p. 219)

In the sections to come, we will explicate in which way this also applies to the DW-DGS. This bilingualized monolingual solution seems to serve the members of the minority SL community as well as learners and other DGS users from the surrounding majority language community.

As already stated, in many cases the lexicographers working on a dictionary of a minority language are not L1 speakers or signers of the object language. The same is true for the DW-DGS, but the project aims to involve the community as much as possible in the compilation process and production of the dictionary. Deaf colleagues and student assistants in the project mainly work on annotation, but also contribute to the lexicographic work (e. g., preparation of examples sentences; questions on the use of a sign; discussion of issues relevant to the community). From early on, the project also involved deaf L1 consultants, the focus group. The lexicographic team of the DW-DGS has regular meetings with the group and discusses topics such as dictionary information types, structures, and layout, or asks for feedback on intermediate stages of the DW-DGS. Thus, the DGS community is involved in various stages of the dictionary-making process.

3.2 Languages in the DW-DGS

With DGS as the object language in focus, information is centered around signs and their use. Object language items are presented with micons (see below) and videos to provide a clear rendering of the form. Object language items are the lemma sign and its forms (variants and modifications), usage examples, collocations and semantic preference patterns, synonyms and antonyms, multi-sign expressions, and cross-references to signs with similar, related, or identical form.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

German is included as second object language to enable bilingual use of the dictionary but is not treated in depth, as it is a well-resourced and documented language. Consequently, we supply translation equivalents for the different senses of a sign, but add little information on the use of these equivalents; instead, we offer a link to the respective entry in the Digitales Wörterbuch der deutschen Sprache (DWDS), an online corpus-based dictionary of German. The DGS examples are also translated into German.

In monolingual dictionaries of widely known languages with writing systems, metalinguistic information normally is given in the same language as the object language. Using a SL for descriptions means resorting to video recordings due to the lack of a functional writing system. However, video recordings are not suitable for every purpose. As a text type, neither the dictionary as a whole nor an entry is read from the beginning to the end, but rather consulted selectively. A presentation of one lengthy video or several shorter video clips would not satisfy the need of the scanning eye. Written language, as a fixed source of information, is needed at least for structural information such as headers of information types (see Kristoffersen/Troelsgård 2012, p. 312), but also for more content-related information. We thus use written German as the language of description.

There is another reason for this decision. Using a minority language for description severely limits the possible range of users. Since the DW-DGS is a hybrid dictionary including bilingual features, the second object language, German, is another option. All headers of information types and signposts for a quick overview of senses but also additional information on use, grammatical notes and the definitions are presented in written German. In using German as the descriptive language, we rely on the bilingualism of deaf users and a larger audience can be reached.

3.3 Sign representation and visual information

Sign language dictionaries need to prioritize visual information for two reasons: the visual nature of their object languages and the visual culture of the target group.

As visual-manual languages, the meaning-distinguishing units in SLs depend on the parameters of movement, handshape, hand orientation, and location. A digital format is thus the ideal choice for a SL dictionary. Signs can then be represented in a direct, detailed, and accurate way through moving images.

As the DW-DGS is an online dictionary, we are able to incorporate videos in our entries. We use videos to show different sign variants and signs that are cross-referenced in the entry (e. g., synonyms). The DW-DGS furthermore includes video examples, taken from the DGS corpus, that show the sign being used in context. As full videos cannot be incorporated in every part of the dictionary, we additionally use micons (moving icons) as small moving representations of signs (see fig. 1).

Fig. 1: Micon of Entry 329

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

The movie thumbnail in the micon shows the sign production when the mouse is hovering over it, giving an impression of the sign's form. Clicking on the movie thumbnail opens the corresponding movie in the movie display area. Hovering over the ID number opens a preview box with the signposts of the sign's senses, giving the user an impression of the sign's meaning. Micons appear in all parts of the dictionary, as they function as small but comprehensive representations of signs.

The second kind of visual information we offer in the dictionary are visualizations of information given in writing. Deaf people, who are part of a visual culture, are a major target group for SL dictionaries. While their functional bilingualism gives them access to the dictionary through written language as well, additional visualizations of certain information types help with accessing that information. We create visualizations in the form of maps showing the regional distribution of signs. Different colors in these maps correspond to different regional signs denoting the same concept (see fig. 2). Geographical signs such as continents, countries and cities are likewise visualized on maps. In this case, the maps are

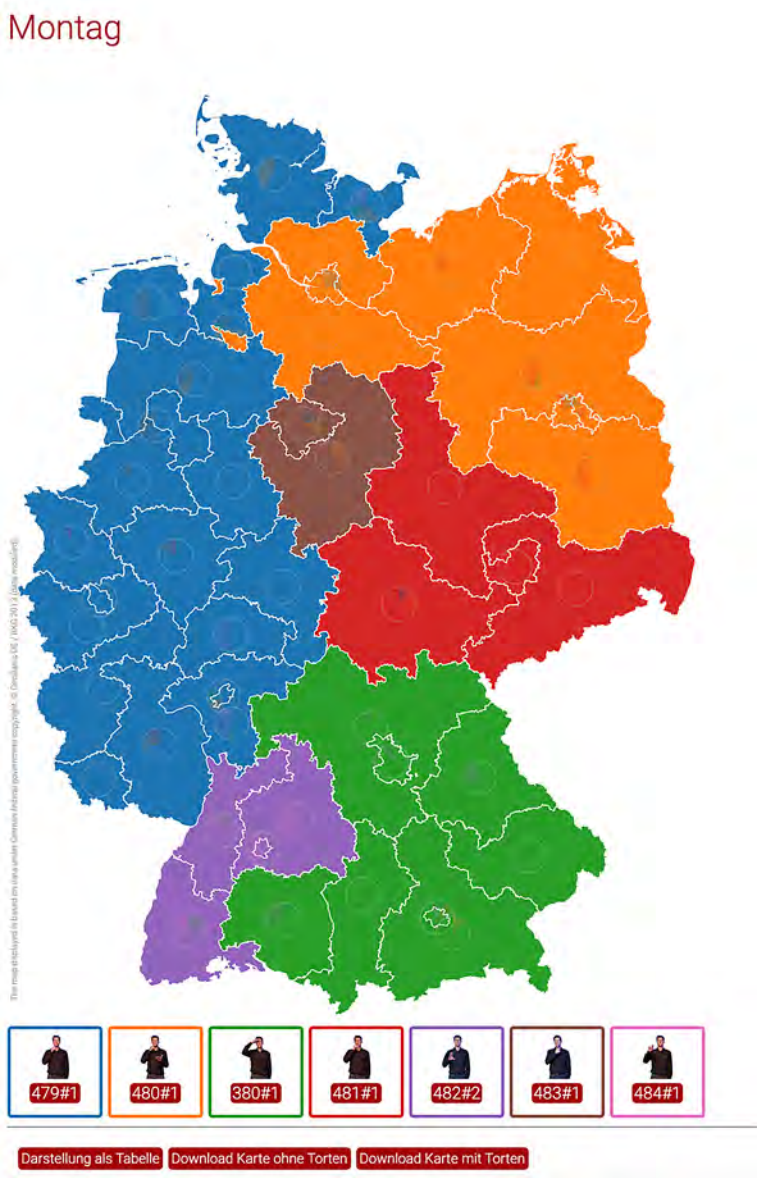


Fig. 2: Map of regional signs denoting 'Monday'

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

geographical maps and the geographical points with entries are represented by dots. Hovering over an area makes the micon of the corresponding sign appear.

We also have a visual access structure to the dictionary in the form of a graph showing the relations that signs have to each other. Entries are represented as dots that are connected by differently colored lines. The colors correspond to different relations such as being the same or similar in form, or containing synonymous senses. The graph is interactive in that different types of relations can be turned on and off and thus invites a playful exploration of the dictionary.

3.4 User groups of the DW-DGS

The DW-DGS is the first corpus-based DGS dictionary and will likely stay the only one of its kind for quite some time. It thus has to serve many user groups at once, a challenge it shares with many other minority language dictionaries. Trying to serve many diverse user groups, a balance has to be found between giving enough information for every user and not giving so much information that it becomes overwhelming for other users. User groups of the DW-DGS include L1 DGS signers, deaf students, L2 learners of DGS, DGS teachers, DGS interpreters, linguists, and the interested public.

Different user groups are served by different information types in the DW-DGS. In table 2, we highlight both monolingual and bilingual uses of the dictionary and state what information types we offer to different users. Neither the use cases nor the information types shown here are meant to be exhaustive.

	dictionary function	**search function**	**relevant information types**
Monolingual			
L1 DGS signer	reception	DGS, form-based	definition, synonyms
	production	DGS, form-based; subject area index	synonyms, variant forms, regional variants
L2 DGS learner (advanced)	reception	DGS, form-based	definition, synonyms, examples
	production	DGS, form-based; subject area index	synonyms, collocations, examples
Bilingual			
L1 DGS signer	reception: German -> DGS translation: German -> DGS	German index	DGS equivalents (signs), synonyms
	production: DGS -> German translation: DGS -> German	DGS, form-based	German equivalents, example translations, links to the DWDS
L2 DGS learner	reception: DGS -> German translation: DGS -> German	DGS, form-based	German equivalents, definition
	production: German -> DGS translation: German -> DGS	German index; subject area index	DGS equivalents (signs), synonyms, examples, collocations

Table 2: Selection of use cases for the DW-DGS

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

To illustrate how individual information types help in specific use cases, we will now describe some of these cases in detail.

The first case could be a deaf person (L1 DGS) who is about to move to a new city. They might want to look up some of the signs used in that region in particular. Weekdays, for example, are highly regional signs in DGS, so this user may use the form-based search function to find their known sign for 'Monday'. In the synonym section of that entry, they would find a link to a map showing all of the regional signs for this concept. They can then move to the entry of the sign used in their new home city. Similarly, they may want to learn the name signs of the cities in that area. They could then use the map of Germany in the appendices which shows all of the geographical name signs described in the dictionary.

A second case could be a hearing person (L2 DGS) with deaf friends who is very interested in sports and would like to sign with their friends about their hobby. They search the subject area index for the topic "sports" look at the signed videos and example sentences in the corresponding entries to learn how to use them. They will not be able to learn all of the relevant signs this way, but they can establish a good basis to start the conversation. This mirrors one of the typical use cases described in Vale (2015) for the NZSL dictionary in which people would look up signs around certain topics in order to prepare for expected communicative situations.

While it is not possible to create a dictionary that is optimized for all user groups at the same time, we nevertheless strive to offer valuable information for each one. Information types like synonyms, collocations, and examples are particularly valuable assets of the DW-DGS, as they provide insight into the sign's meaning in DGS and help differentiate the senses of a sign. From that perspective, they are ideal for L1 users. However, they are just as helpful for L2 users, albeit in different ways. These information types provide insights into contexts of the actual usage of a sign and help extend the learner's vocabulary. Thanks to our data basis, the DGS corpus, we are able to give valuable information in this area. The other side of the coin is that the dictionary is not optimized for the German-to-DGS direction of use. The German index is simply a collection of the translational equivalents given in each sign entry, that is, it is not systematically built and incomplete with regard to included words and their senses.

4. Conclusion

Building a dictionary for a sign language comes with certain challenges: some based on the aspect of it being a minority language (lack of resources, many user groups), others based on the specific community structure of the L1 community (visual culture, functional bilingualism), and again others based on the visual-manual modality of the language (lack of writing system, visual representation). We have described how we tackle each of these challenges in the DW-DGS by compiling a corpus-based dictionary that is fit for many user groups due to using both DGS and German, exact sign representations, and visualized information.

The future will show how our dictionary is used by different user groups and in how far our efforts prove successful. As of now, preliminary entries of the dictionary are available online, allowing us to get feedback from the language community and the general public. Reactions have been positive, though they have also shown that there is in fact a learning curve on part of the users, who are not used to navigating such dictionary structures. There

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

is however a lot of interest in the contents of the dictionary and L1 signers especially are excited to see their language described. We are happy to be able to get feedback on these matters during the compilation of the dictionary and look forward to seeing the reception of the final product.

References

Adamska-Sałaciak, A. (2013): Issues in compiling bilingual dictionaries. In: Jackson, H. (ed.): The Bloomsbury companion to lexicography. London, pp. 213–231.

Atkins, B. T. S./Rundell, M. (2008): The Oxford guide to practical lexicography. Oxford/New York.

BGG (2002): Gesetz zur Gleichstellung von Menschen mit Behinderungen. https://www.gesetze-im-internet.de/bgg/index.html (last access: 09-05-2022).

Bradley, D. (2015): The lexicography of minority languages in Southeast Asia. In: Hanks, P./de Schryver, G.-M. (eds.): International handbook of modern lexis and lexicography. Berlin/Heidelberg, pp. 1–11. https://doi.org/10.1007/978-3-642-45369-4_97-1.

(CDPSL) Corpus-based Dictionary of Polish Sign Language (2016). http://www.slownikpjm.uw.edu.pl/en (last access: 18-05-2022).

Cristinoi, A./Nemo, F. (2013): Challenges in endangered language lexicography. Lexicography and dictionaries in the information Age. https://halshs.archives-ouvertes.fr/halshs-01345620.

De Meulder, M./Murray, J. J./McKee, R. L. (2019): The legal recognition of sign languages: advocacy and outcomes around the world. Bristol. https://doi.org/10.21832/9781788924016.

DWDS – Digitales Wörterbuch der deutschen Sprache (2007–): https://www.dwds.de/ (last access: 09-05-2022).

Dorian, N. C. (2006): Minority and endangered languages. In: Bhatia, T. K./Ritchie, W. C. (eds.): The handbook of bilingualism. Oxford, pp. 437–459. https://doi.org/10.1002/9780470756997.ch17.

Erlenkamp, S. (1998): Lexikographie als Teil einer Minderheitenpolitik – Minderheitenpolitik als Teil einer Lexikographie. In: Das Zeichen 43, pp. 98–105.

Hall, M. L./Hall, W. C./Caselli, N. K. (2019): Deaf children need language, not (just) speech. In: First Language 39 (4), pp. 367–395. https://doi.org/10.1177/0142723719834102.

Harris, M./Terlektsi, E. (2021): Literacy attainment among children who are deaf or hard of hearing: The past, the present, and the future. In: Easterbrooks, S. R./Dostal, H. M. (eds.): The Oxford handbook of deaf studies in literacy, pp. 9–23. http://doi.org/10.1093/oxfordhb/9780197508268.013.2.

Hartman, M. C./Nicolarakis, O. D./Wang, Y. (2019): Language and literacy: issues and considerations. In: Education Sciences 9 (3), 180, pp. 367–395. https://doi.org/10.3390/educsci9030180.

Hilzensauer, M./ Bergmeister, E./Dotter, F./Krammer, K./Okorn, I./Orter, R./Skant, A. (2000): Zum Stand der Forschung in der Gebärdensprachlexikographie. In: Das Zeichen 52, pp. 288–305.

Johnston, T. (1998): Signs of Australia. A new dictionary of Auslan (the Sign Language of the Australian Deaf Community). Revised edition. North Rocks, NSW.

Kristoffersen, J. H./Troelsgård, T. (2012a): Integrating corpora and dictionaries: problems and perspectives, with particular respect to the treatment of sign language. 5th Workshop on the Representation and Processing of Sign Languages (LREC 2012), 2012-05-27, Istanbul, Turkey. [presentation slides].

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Kristoffersen, J. H./Troelsgård, T. (2012b): The electronic lexicographical treatment of sign languages: The Danish sign language dictionary. In: Granger, S./Paquot, M. (eds.): Electronic lexicography. Oxford, pp. 293–315.

Langer, G./Müller, A./Wähl, S. (2018): Queries and views in iLex to support corpus-based lexicographic work on German sign language (DGS). In: Bono, M./Efthimiou, E./Fotinea, S./Hanke, T./Hochgesang, J./Kristoffersen, J./Mesch, J./Osugi, Y. (eds.): Proceedings of the LREC2018 8th Workshop on the Representation and Processing of Sign Languages: Involving the Language Community. 11th International Conference on Language Resources and Evaluation (LREC 2018), Miyazaki, Japan, 12.5.2018. Paris, pp. 107–114.

Langer, G./Schulder, M. (2020): Collocations in sign language lexicography: towards semantic abstractions for word sense discrimination. In: Efthimiou, E./Fotinea, S./Hanke, T./Hochgesang, J./Kristoffersen, J./Mesch, J. (eds.): Proceedings of the LREC2020 9th Workshop on the Representation and Processing of Sign Languages: Sign Language Resources in the Service of the Language Community, Technological Challenges and Application Perspectives. 12th International Conference on Language Resources and Evaluation (LREC 2020), Marseille, Frankreich, 16.5.2020. Paris, pp. 127–134.

Maxwell, M. (1985): Some functions and uses of literacy in the deaf community. In: Language in Society 14 (2), pp. 205–221. https://doi.org/10.1017/S0047404500011131.

McBurney, S. L. (2006): Sign language: history of research. In: Brown, K./Anderson, A. (eds.): Encyclopedia of language/linguistics. 2nd edition. Amsterdam/Boston, pp. 310–318. https://doi.org/10.1016/B0-08-044854-2/01332-8.

McKee, R. L./McKee, D. (2017): The online dictionary of New Zealand sign language: a case study of contemporary sign language lexicography. In: Fuertes-Olivera, P. A. (ed.): The Routledge handbook of lexicography. London, pp. 399–420.

Mosel, U. (2004): Dictionary making in endangered language communities. In: Austin, P. K. (ed.): Language documentation and description. Vol. 2. London, pp. 39–54. http://www.elpublishing.org/docs/1/02/ldd02_04.pdf

Moskovitz, D. (1994): The dictionary of New Zealand sign language user requirements survey. In: Ahlgren, I./Bergman, B./Brennan, M. (eds.): Perspectives on sign language: papers from the Fifth International Symposium on Sign Language research, Salamanca, Spain, 25–30 May 1992. Vol. 2: Perspectives on Sign Language usage. Durham, pp. 421–442.

(ODT) Ordbog over Dansk Tegnsprog. Center for Tegnsprog (2008–2016): http://www.tegnsprok.dk (last access: 25-03-2022).

Paul, P. V. (2018): Literacy, literate thought, and deafness. In: American Annals of the Deaf 163 (1), pp. 78–86. https://doi.org/10.1353/aad.2018.0013.

Plaza-Pust, C. (2016): Bilingualism and deafness: on language contact in the bilingual acquisition of sign language and written language. Boston/Berlin/Preston.

Power, M. R./Power, D. (2004): Everyone here speaks TXT: deaf people using SMS in Australia and the rest of the world. In: Journal of Deaf Studies and Deaf Education 9 (3), pp. 333–343.

Prinsloo, D. J. (2012): Electronic lexicography for lesser-resourced languages: the South African context. In: Granger, S./Paquot, M. (eds.): Electronic lexicography. Oxford, pp. 119–143. https://doi.org/10.1093/acprof:oso/9780199654864.003.0007.

Stokoe, W. C./Casterline, D. C./Croneberg, C. G. (1965): A dictionary of American sign language on linguistic principles. Washington, DC.

Stokoe, W. C. (1993): Dictionary making, then and now. In: Sign Language Studies 1079 (1), pp. 127–146. https://doi.org/10.1353/sls.1993.0017.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Vale, M. (2015): A study of the users of an online sign language dictionary. Electronic lexicography in the 21st century: Linking lexical data in the digital age. In: Kosem, I./Jakubíček, M./Kallas, J./Krek, S. (eds.): Proceedings of the ELex 2015 Conference, 11–13 August 2015, Herstmonceux Castle, United Kingdom. Ljubljana/Brighton, pp. 281–303.

Wright, S. (2018): Planning minority language maintenance: challenges and limitations. In: Rehg, K. L./Campbell, L. (eds.): The Oxford handbook of endangered languages. New York, pp. 636–656. https://doi.org/10.1093/oxfordhb/9780190610029.013.30.

Zwitserlood, I./Kristoffersen, H./Troelsgård, T. (2013): Issues in sign language lexicography. In: Jackson, H. (ed.): The Bloomsbury companion to lexicography. London, pp. 259–283.

Contact Information

Anke Müller
University of Hamburg
anke.mueller@uni-hamburg.de

Gabriele Langer
University of Hamburg
gabriele.langer@uni-hamburg.de

Felicitas Otte
University of Hamburg
Felicitas.Otte@uni-hamburg.de

Sabrina Wähl
University of Hamburg
sabrina.waehl@uni-hamburg.de

Acknowledgements

This publication has been produced in the context of the joint research funding of the German Federal Government and Federal States in the Academies' Program, with funding from the Federal Ministry of Education and Research and the Free and Hanseatic City of Hamburg. The Academies' Program is coordinated by the Union of the German Academies of Sciences and Humanities.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Specialized Dictionaries

XX EURALEX INTERNATIONAL CONGRESS
DICTIONARIES AND SOCIETY
12-16 July 2022, Mannheim, Germany

Maria Aldea

BIEN ÉCRIRE, BIEN PARLER AU XIXE SIÈCLE. LE RÔLE DU DICTIONNAIRE DANS L'APPRENTISSAGE DE LA LANGUE MATERNELLE: LE CAS DU ROUMAIN

Abstract In this paper, the author studies the role of the dictionary in the first language acquisition, highlighting its didactic value. Based on two Romanian lexicographical works of the 19th century, *Lexiconul de la Buda* (Buda, 1825) [*the Lexicon of Buda*] et *Vocabularu romano-francesu* (Bucarest, 1870) [*the Romanian-French Vocabulary*], the author analyses the normative information recorded in the articles in order to observe which level of language (i. e. phonetical, morphological, syntactical and lexical) is concerned. Such an approach allows to distinguish between the possible changings both at the level of the perception or at the grammatical, lexical and semantical description, i. e. the settlement of the word in the first language, and at a technical level, i. e. the making of article and of dictionary.

Keywords First language acquisition; dictionary; linguistic norm; cultural norm; Lexiconul de la Buda; Vocabularu romano-francesu

1. Introduction

Depuis toujours, les ouvrages à finalité didactique tels que les manuels scolaires, les grammaires et les dictionnaires ont joué un rôle essentiel dans l'apprentissage d'une langue, soit-elle maternelle ou étrangère, tout en permettant aux jeunes apprenants de s'approprier la langue et de l'assimiler dans sa variante soignée aussi bien au niveau écrit qu'au niveau oral. L'acquisition des outils langagiers corrects imposés par une expression soignée dans une langue donnée (n'importe l'époque) révèle, en fin de compte, la volonté de tout apprenant de bien écrire dans cette langue et de bien la parler. Se déplaçant entre ces deux pôles, celui de la langue commune et celui de la langue soutenue ou littéraire, le jeune apprenant s'efforcera d'acquérir et d'employer plutôt la seconde, car ce sera celle-ci qui lui conférera la marque distinctive d'appartenance à une culture intellectuelle supérieure, à un milieu prestigieux.

C'est dans une telle perspective qu'on verra s'inscrire aussi le processus d'acquisition de la langue roumaine. Ayant connu un parcours historique plutôt sinueux et ayant subi, selon les époques, des influences diverses venues de différents mouvements ou de cultures étrangères qui ont mis leur empreinte sur l'essor ou la stagnation du roumain (tant au niveau administratif qu'au niveau éducatif ou culturel), la langue roumaine a réussi à s'imposer assez tardivement en tant que langue officielle de l'administration et en tant que langue de culture, en remplaçant le slavon vers la fin du XVIIIe siècle (cf. Aldea 2018a). D'ailleurs, elle restera tributaire à une tradition graphique et à une écriture à caractères cyrilliques pendant plus de trois siècles. Néanmoins, la fin du XVIIIe siècle marquera un tournant dans ce sens, dû surtout aux efforts intellectuels menés par plusieurs érudits de l'École latiniste de Transylvanie afin de remplacer les lettres cyrilliques par des caractères latins et par une orthographe étymologiste. Le but recherché était de rendre visible « le génie de la langue roumaine » et de redonner à celle-ci ses lettres de noblesse en tant que langue descendant directement du latin comme ses autres langues-sœurs romanes (ibidem).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Ce passage vers une écriture à caractères latins et vers une orthographe étymologiste connaîtra des états différents selon les trois provinces historiques roumaines – la Transylvanie, la Moldavie et la Valachie – où le phénomène se produit. En Transylvanie, ce processus se fera sentir plus fortement à partir des deux dernières décennies du XVIIIe siècle; la publication croissante de livres et de manuels scolaires traduits en roumain et rendus à l'aide de caractères latins et d'une orthographe étymologiste, de même que les tentatives d'élaborer des dictionnaires ayant comme langue de base le roumain vont contribuer à l'uniformisation progressive et à la normalisation graduelle de la langue roumaine (voir Gheţie 1975; Costinescu 1979; Gafton 2012; Chivu 2019, p. 143–208; Aldea 2018a, et d'autres). Jusqu'à la fondation de l'Académie roumaine en 1866 dont les objectifs seront, entre autres, de fixer la langue par la publication d'une grammaire et d'un dictionnaire du type « Trésor » de la langue roumaine, de nombreux produits livresques reflètent l'évolution des modes intellectuelles et les goûts littéraires de leurs auteurs (d'ailleurs, un grand nombre de textes traduits, mais aussi de textes originaux vont paraître à cette époque).

De ce vaste ensemble d'ouvrages à finalité didactique (voir Ghibu 1998), nous avons choisi de nous pencher dans ce qui suit sur les dictionnaires, et plus particulièrement sur leur rôle dans l'apprentissage de la langue maternelle, étant donnée surtout la fonction prescriptive remplie par les dictionnaires qui, comme les grammaires, ont le rôle de fixer une norme et d'imposer un usage correct de la langue (voir J. Dubois 1970, p. 35–47; J. Dubois et Cl. Dubois 1971, p. 49–56, 99–104; Zgusta 1971, p. 164–196; Svensén 1993, p. 44–48; Cabré 1998, p. 237–251; Atkins et Rundell 2008, et d'autres). Dans ce sens, notre attention sera fixée notamment sur les entrées présentant dans leur description des notices supplémentaires d'ordre normatif, car ces notices reflètent non seulement la norme de la langue littéraire, mais aussi la langue vivante et parlée, tout en nous offrant une image particulière sur la langue, sur l'époque dans laquelle elle s'inscrit, de même que sur la personnalité des rédacteurs, ces artisans des goûts culturels. Ainsi, notre corpus se fondera sur deux dictionnaires roumains du XIXe siècle, intitulés *Lexiconul de la Buda* (Buda, 1825) [*le Lexicon de Buda*] et *Vocabularu romano-francesu* (Bucarest, 1870) [*le Vocabulaire roumain-français*] (désormais abrégés LBe, respectivement VRF) (voir Seche 1966).

2. Analyse du corpus

Afin de mieux observer le rôle que ces deux dictionnaires ont joué dans l'apprentissage de la langue maternelle au XIXe siècle, nous nous proposons dans un premier temps d'entreprendre une sélection des entrées enregistrées tant par le LBe que par le VRF, présentant dans leur description des notices d'ordre normatif. Ainsi, nous avons sélectionné environ 2200 articles de dictionnaire dont environ 400 ont été puisés dans le LBe tandis que le reste des exemples, réunissant environ 1800 articles, ont été repris du VRF.

L'examen de notre corpus met en évidence un écart considérable entre l'architecture de nos deux dictionnaires, *id est* la nomenclature, le choix de l'orthographe étymologiste ou la manière de signaler les différentes indications normatives dans la microstructure de l'article. Dans ce qui suit, nous nous proposons de retracer quelques-unes de ces particularités.

En ce qui concerne la nomenclature et le corps proprement dit de l'article correspondant au roumain, nous constatons que, par rapport au VRF qui est rédigé exclusivement en lettres latines et en orthographe étymologiste, le LBe emploie pour le mot-titre des caractères latins et une orthographe étymologiste tout en le doublant par son équivalent graphique rendu en

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

lettres cyrilliques. En effet, ce geste permettra aux lecteurs/apprenants de mieux s'approprier et aux auteurs de leur faciliter la compréhension de ce nouveau type d'écriture en lettres latines à propos de laquelle nous ne pouvons pas encore parler en termes de tradition, mais de tout au plus quelques tentatives d'implémentation. Pour la séquence en roumain du corps de l'article, les rédacteurs emploient seulement des lettres latines et une orthographe étymologiste. Bien qu'on puisse identifier des concordances graphiques dans les deux ouvrages (comme, par exemple, la présence des consonnes géminées) et malgré l'usage de l'orthographe étymologiste, on constate néanmoins des différences graphiques au niveau de la restitution du même son. Par exemple, pour rendre le son /ă/ les rédacteurs du LBe emploient, selon l'origine du mot, les voyelles *a*, *e*, *i*, *o*, *u* individualisées par une petite virgule au-dessus de chacune d'entre elles, de sorte qu'on y voie *ả*, *ẻ*, *ỉ*, *ỏ*, *ủ* (rapprochant, de cette façon, les mots de leurs étymons) (citons, en guise d'exemples: *açiả*[1] /atsə/ 'fil'; *adequỏ* /adekə/ 'c'est-à-dire'; *bătủtură* /bətəturə/ 'trame, duite'; *vỉduvă* /vəduvə/ 'veuve'; *zẻbăvire* /zəbəvire/ 'attardement'; s. v.). De son côté, le rédacteur du VRF emploie dans le premier tome le graphème *â* (par exemple: *adressâ* /adresə/ 'adresse'; *împroșcâturâ* /ɨmproʃkəturə/ 'flaqué,-e'; s. v.) et dans le second tome le graphème qu'on utilise encore de nos jours, *ă* (par exemple: *ligatură* /ligaturə/ 'ligature'; *pardossellă*, *pardossire* /pardosealə/ /pardosire/ 'plancher'; s. v.). Et les exemples pourront bel et bien continuer.

Pour ce qui est effectivement des notices normatives, nous remarquons plutôt l'absence du caractère méthodique ou systématique de l'acte de les individualiser. Ainsi, dans le LBe elles se présentent soit en roumain, soit en latin; elles peuvent ou non être signalées par des formules du type *Nota* 'bien noter que' rendu de différentes façons comme, par exemple, *Nota*, *Not.*, *NB.* (voir les exemples de 1 à 5) ou *Usu* 'usage' rendu par *Usu* et *Us.* (voir les exemples 6 et 7); enfin, elles peuvent se montrer dans certains cas sous la forme de séquences plus ou moins délimitées par des parenthèses rondes (voir l'exemple 8). Dans la plupart des situations ces notices se placent à la fin de la description (voir les exemples 2, 3, 5, 9), mais on les trouve également au début du corps de l'article (voir l'exemple 6) ou au milieu de celui-ci (voir les exemples 1, 4, 7, 8) fonctionnant au sein d'un commentaire définitionnel ou en tant qu'information supplémentaire.

(1) **Anu** 'année', *m. pl.* ani. […]. *Not. subin subst. hoc summitur pro adverbio temporis pro* acum un an, seau acum au trecut anu 'il y a un an ou un an est déjà passé' […].

(2) **Babả** 'vieille femme; guérisseuse', *f. pl.* be. […]. *Not.* în unele părți a Țărei ungurescă se întrebuințează coventu acest în loc de Mamă 'dans certaines parties du Pays hongrois on emploie ce mot à la place du mot *mère'*. *Ital.* Babbo. *i. e.* Pater.

(3) **Departe** 'loin' […]. *I. adj.* […]. *II. adv.* […]. *Nota.* se întrebuințează și în loc de substantiv. *V.* Depărtare, Nro. 2 'on l'emploie également à la place du nom. V. Éloignement, nro 2'.

(4) **Sẻne** 'soi-même' […]. – *(Nota: hic sumitur pro substantivo) propria sponte* […].

(5) **Caieru** 'quenouillée' *m. pl.* re. *f.* […]. NB. 1) *Caliendrum denotat comam adsciticiam, et hinc per analogiam*, Caieru. […].

(6) **Chilinescu**, seau clinescu, magis in usu: desclinescu, -nire, -nitu 'distinguer'. […]. *verb. act.* […].

[1] Pour une identification plus rapide du mot vedette dans la nomenclature, nous le rendons avec la forme graphique avec laquelle il est consigné dans les deux dictionnaires. Exceptant les exemples de 1 à 16, pour des contraintes d'ordre typographique, nous choisissons d'indiquer seulement la traduction française du contenu de ces notices normatives.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

(7) **Harmoníe**, seau armoníe 'harmonie', *f. pl.* ii. [...]. *Us.* cânt la, seau cu harmonie 'je chante en ou avec harmonie' [...].

(8) **Cicnescu, nire, nitu** 'mourir'. [...]. *verb. act.* [...] (se zice despre făpturile ceale necuvântătoare 'on le dit à propos des créatures qui ne parlent pas') [...].

(9) **Inde** 'où'. [...]. *adv.* zic unii în loc de unde, *V.* Unde 'certains l'emploient à la place du mot *où*, V. Où'.

En revanche, dans le VRF, exceptant quelques entrées (voir l'exemple 10), ces indications normatives sont consignées uniquement en roumain, pouvant ou non être encadrées par des parenthèses rondes (voir les exemples de 11 à 16). Au sein de l'article, ces indications sont placées soit au début de la définition (voir les exemples 11, 12, et 14), soit à la fin de celle-ci (voir les exemples 13, 15 et 16), en complétant ainsi le commentaire définitionnel.

(10) **Dadâ** 'dada, grande sœur'. *s. f.* Lele, leica, leliţe. Titlu ce dau copii<i>, şi după copii şi alţii, sorii sau şi altei femei mai mare 'Tante, tatie. Titre donné par des enfants et, dans le sillage des enfants, par d'autres personnes aussi, à une sœur ou à une autre femme plus âgée'. Dada. (*Explication de Ménage.*)

(11) **Cépâ** 'oignon'. *s. f.* (se cetește ceapă 'se lit oignon') [...].

(12) **Desertŭ** 'désert'. *s.etr.* (Dezert) [...].

(13) **Mirare** 's'émerveiller, s'étonner', *v. s.* [...] (se conjugă, cu pasivele *me*, *te*, *se*, *ne*, *vă*, *se*, ca *laudare* 'il se conjugue avec les pronoms *me*, *te*, *se*, *nous*, *vous*, *se*, comme *(se) vanter*').

(14) **Muzică** 'musique'. (vezi Musică 'voir Musique').

(15) **Cĭocoiŭ** '*exploiteur*'. *s. m.* [...]. – (Ţăranii poreclesc asfel pe toţi funcţionarii, pe toţi boierii 'titre donné par les paysans aux employés ou aux boyards').

(16) **Palo-de-vaca**. *s. m. t. de bot.* Numele unui arbure care dă un lapte foarte bun, aflat în Venezuela de savantul Humbold. Se zice şi arburele vacei sau arbure de lapte 'Le nom d'un arbre qui donne un très bon lait, découvert au Vénézuela par le savant Humboldt. On l'appelle également l'arbre de la vache ou l'arbre à lait'. *Palo-de-vaca.*

2.1 Type d'information et niveau de langue

Dans ce qui suit, nous nous proposons d'analyser le type d'informations (*i. e.* phonétique, morphologique, syntaxique et/ou lexicale) contenues dans ces notices normatives.[2]

2.1.1 Niveau phonétique

Dans le LBe, les rédacteurs tiennent à apporter des indications concernant la prononciation de différents mots en langue maternelle, tout en recourant parfois à une perspective comparée avec d'autres langues romanes ou avec d'autres dialectes du roumain.

Ainsi, nous notons pour le roumain que le mot à valeur interjectionnelle *ho !* 'halte-là !' suppose « une prononciation longue » (s. v. *Ho!*), tandis que les noms *ţucăr* et *zăhar* 'sucre' sont à prononcer « avec une prononciation rude » (s. v. *Mére de trestie* 'miel de canne').

[2] Dans cette étude nous ne discuterons pas les notices introduisant des remarques étymologiques ou culturelles.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Pour le dialecte aroumain employé au sud du Danube, les rédacteurs attirent l'attention sur le correspondant aroumain du mot roumain *dar* sau *dară* 'mais', soulignant que « *Valachi Aurelianae Daciae dicunt*: decare, *hinc est*, dară, *contractum* » (s. v. *Dar'*, séu *dare*). Dans le même sens, pour d'autres communautés linguistiques, ils avancent différentes remarques telles que « *Florentini cacabus pronunciant hahabus* » (s. v. *Cåhalå*), ou « *apud Hispanos*: *cha, che, chi, cho, chu, pronunciantur sicut in Toscana cia, ce, ci, cio, ciu. Valachi Hispanico*: chiste, *interponunt n more solito, et fit*: cinste, cinstire » (s. v. *Cinstire*).

Nous avons pu également répertorier plusieurs notices concernant l'aspect graphique. Ainsi, le mot roumain *inimă* 'cœur' « est censé s'écrire de la manière suivante, à savoir *'animå*, mais il faut d'habitude le chercher dans la section affectée à la lettre I, c'est-à-dire *inima* 'cœur' » (s. v. *'Animå*). Dans le cas du mot roumain *papistaş* 'catholique', on recommande de « l'écrire *catholique romain* » (s. v. *Papistaşu*); il est également indiqué de rendre par « deux mots » le nom *întâiul-născut* 'le premier-né' (s. v. *'Antẻiu nåscutu*). Malgré les trois variantes graphiques retenues en position de mot vedette, *i. e. halap, ialap* et *jalapă* 'jalap', les rédacteurs du LBe introduisent dans le corps de l'article une indication supplémentaire: « *Nota.* Cette racine d'Inde s'appelle aussi *gialapă* 'jalap': *gialappa, Schroed* » (s. v. *Halapu, séu Ialapu, m. pl. i. séu jalapå*).

En revanche, dans le VRF, les indications concernant la prononciation et l'orthographe d'un mot sont soit signalées dans le corps de l'article à l'aide d'informations placées entre parenthèses rondes, rarement renforcées par la présence du verbe métalinguistique « *se citeşte* » 'il se lit', soit effectivement intégrées dans la définition même du mot. Par exemple, le pronom démonstratif féminin rendu graphiquement par la forme *aquea* 'celle' renferme dans sa notice même une information concernant sa prononciation (*Il se lit acea* 'celle') » (s. v. *Aquea şi A*). Dans d'autres cas, comme, par exemple, « *Această-l-altă*: Celle-ci. *Aceste-l-alte*: Celles-ci », le rédacteur avance des explications plus précises: « (Entre ces démonstratifs et les indéfinis qui les composent on intercale parfois un *-l-* pour des raisons d'euphonie, comme on peut le voir.) » (s. v. *Aquéstâ şi Astâ*); voir aussi l'entrée consacrée à l'adverbe *alaltăieri* 'avant-hier', prononcé « (A-l-altă-eri) » (s. v. *Alaltaerĭ*), ou celle consacrée au mot *armăsar* 'étalon' où l'on indique que le « (*H se prononce dans la gorge*) » (s. v. *Harmâssarŭ, Armâssarŭ*), etc. À ces exemples, on peut en ajouter encore deux, dont la forme orale est simplement marquée entre parenthèses: le nom *desertŭ* 'désert' qui se prononce « (Dezert) » (s. v. *Desertŭ*) et l'adjectif *exemplariu* 'exemplaire' qui se prononce « (*ekzemplarĭu*) » (s. v. *Exemplarĭu - â*).

En examinant l'inventaire de tous ces mots bénéficiant de telles remarques normatives d'ordre phonétique, nous avons constaté que la plupart d'entre eux visaient surtout la catégorie des mots perçus à ce moment-là comme étant des néologismes. De plus, nous avons sélectionné aussi des articles qui contenaient des notices traitant des variantes ou des doublets graphiques; nous en retenons ici deux exemples: pour le verbe *a lăţui* 'latter', grâce à la notice placée à la fin de l'article, nous constatons que le rédacteur attire l'attention sur le doublet graphique du mot vedette tout en l'expliquant: « (les uns écrivent *lănţuire*, peut-être à cause de l'arrangement des planches de bois qui s'entassent les unes sur les autres pareilles aux maillons d'une chaîne) » (s. v. *Lăţuire*). Par contre, pour le nom *blusă* 'blouse', la notice normative précède la définition, tout en mettant en évidence que « (les uns écrivent *bluză* 'blouse') », cette dernière forme graphique étant, d'ailleurs, celle employée de nos jours (s. v. *Blusâ*). Dans d'autres cas, le rédacteur remarque l'existence de confusions phonétiques et/ou graphiques, des confusions qu'il justifie par la méconnaissance du sens des mots en question par les parleurs roumains. Mentionnons, en guise d'exemples, les mots *meară* 'pomme' et

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

măr 'pommier' qui sont prononcés et écrits « *mărŭ, Mêrŭ* (mot qui renvoie à l'arbre et non pas au fruit), c'est-à-dire le fruit rond et bon à manger du pommier » (s. v. *Méră*). La liste de tels exemples reste, d'ailleurs, plutôt riche (voir Aldea 2018b).

2.1.2 Niveau morphologique et syntaxique

Pour ce qui est des aspects morphologiques et syntaxiques, les remarques des rédacteurs portent, en général, sur l'usage des parties du discours, les désinences de pluriel, la conjugaison des verbes et l'ordre de mots.

Les rédacteurs du LBe signalent plusieurs situations dans lesquelles le même mot peut acquérir des valeurs morphologiques différentes dans des contextes distincts. Par exemple, l'adverbe *adecă* 'c'est-à-dire' « *adv. hoc subin usurpatur pro substantivo, et tunc significant finis, exitus* » (s. v. *Adequȯ, s. adechê*), le verbe à valeur neutre *a custa* 'vivre' « *verbum hoc neutrum usurpatur in usu communi pro activo* » (s. v. *Custu, stare, statu*), l'adverbe *iacă* ou *iacătă* 'voici, voilà' (apprécié de nos jours en tant qu'interjection) « s'emploie avec le pronom personnel *je, tu, il, nous, vous, ils* » (s. v. *Éccê, séu éccête*); enfin, l'emploi identique dans des contextes différents du mot *feliu* 'mode, manière' est signalé par la notice: « *Nota. a*) Tous ces exemples relatifs à l'usage de ce mot en tant qu'adjectif, remplaçant la préposition *de* 'de' par la préposition *în* 'dans', ne poseront pas de problèmes si l'on respecte les normes grammaticales » (s. v. *Féliu*), etc.

Nous remarquons également la présence de plusieurs notices concernant l'emploi de certains mots perçus à l'époque comme des prépositions, dont les uns conservent encore ce statut jusqu'à nos jours tandis que d'autres se retrouvent aujourd'hui soit parmi les adverbes, soit à l'intérieur de la classe des affixes lexicaux: la préposition *poi* 'après' (qui est à retrouver de nos jours seulement comme adverbe) « ne s'emploie pas toute seule dans un énoncé, mais toujours en combinaison avec d'autres mots » (s. v. *Poĭ*); pour ce qui est de la préposition *de* 'de', les rédacteurs notent qu'« [i]l y a beaucoup de situations où l'on peut employer la particule *de* mais son emploi en combinaison avec une partie du discours ou une autre (*partium orationis*) est justifié par le contexte; quelques-uns de ces emplois sont à retrouver dans le présent Lexicon, soit ici, soit dans d'autres endroits; pour d'autres utilisations, il faudra se fier à l'emploi commun et à la *syntaxis verborum* » (s. v. *De*); enfin, en ce qui concerne le mot *strå* 'avant', « il ne s'emploie jamais seul, mais toujours en combinatoire avec d'autres parties du discours (*partibus orationis*), comme ci-dessous » (s. v. *Strå*).

Bien que la plupart des noms et des adjectifs présentent un encadrement lexico-grammatical et la mention des terminaisons renvoyant au pluriel (dans le cas des noms) et au genre et au nombre (dans le cas des adjectifs), nous avons pu identifier des entrées où les rédacteurs ont ressenti le besoin de préciser ces aspects une fois de plus dans le corps de l'article, tout en apportant des explications relatives à leur sémantisme [par exemple, « *Nota.* Par grains *plur.*, on entend parfois des céréales » (s. v. *Gråuntiu* 'Grain'); ou, quand le mot *lăptucă* 'laitue' signifie *salade*, il « s'emploie surtout au pluriel, c'est-à-dire *lăptuci* » (s. v. *Låptucå*)] ou à leur usage le plus fréquent [*Nota.* Le mot *grumaz* 'cou' « s'emploie plutôt au pluriel » (s. v. *Grumazu*)]. Il y a des cas où l'on indique la forme du pluriel soit dans le corps du même article [le nom *obeadă* 'jante' « *pluralem format*: obeade 'jantes' » (s. v. *Obédå. f.*)], soit dans une entrée à part [« Grâne, seau Grâneaţe 'céréales', *f. plur. ex* grâu seau grânu 'blé ou céréales' » (s. v. *Gråne, séu Grånéţe*); « Oameni 'hommes' […], *plur. ex* om 'homme' » (s. v. *Ómeni*)].

Nous attirons également l'attention sur la présence de remarques concernant le fonctionnement syntaxique de certaines constructions roumaines dans les autres langues employées

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

dans le dictionnaire. Retenons à titre d'exemple le fonctionnement distinctif de la structure roumaine « [scap] pre cineva fără voia mea » 'je sauve quelqu'un malgré moi' tandis que « dans les langues latine et allemande (ou saxonne), la structure est entièrement renversée » (s. v. *Scapu, scâpare, patu*).

En ce qui concerne les notices d'ordre morphologique et syntaxique consignées dans le VRF, on remarque la présence accrue des indications relatives à la conjugaison des verbes et précisées par la formule « se conjugă ca » 'il se conjugue comme' + l'infinitif du verbe prototype, formule placée d'habitude à la fin de l'article. On signale, ainsi, le fait que le verbe *a atenta* 'attenter' « se conjugue comme *lucrare* 'travailler' » (s. v. *Attentare*), le verbe *a cicăli* 'agacer, harceler' « se conjugue comme *gîndire* 'penser' » (s. v. *Cicălire*), etc., tandis que d'autres verbes bénéficient dans leur description d'une palette assez large d'occurrences verbales aussi bien à l'indicatif présent qu'à tous les modes du verbe. Prenons en guise d'exemple le verbe *a continua* 'continuer' qui « [s]e conjugue comme *lăudare* 'vanter, glorifier'. *Continuu* 'je continue', *continui* 'tu continues', *continuă* 'il continue', *continuăm* 'nous continuons', *continuaţi* 'vous continuez', *continuă* 'ils continuent', *etc.* » (s. v. *Continuare*).

Rarement, à ces précisions normatives s'ajoutent aussi des indications concernant différents usages ou leurs encadrements grammaticaux: le verbe *a defida* 'défier' « [s]e conjugue comme *lăudare* 'vanter, glorifier'. Certaines personnes le considèrent comme appartenant au troisième groupe de conjugaison – *defidere* – et le conjuguent comme *înţelegere* 'comprendre') » (s. v. *Defidare*).

De nombreux articles contiennent dans leur description des précisions (qui ne sont pourtant pas signalées par des parenthèses rondes) relatives aux valeurs morphologiques ou à la fonction syntaxique qu'un mot pourrait acquérir dans un contexte donné. Ainsi, l'adverbe *da* 'oui' « s'emploie aussi en tant que nom » (s. v. *Da*), la conjonction conditionnelle *dacă* 'si' « s'emploie parfois comme nom » (s. v. *Daca*) et le nom masculin *decagon* 'décagone' « s'emploie aussi en tant qu'adjectif » (s. v. *Decagonŭ*). Dans le cas du verbe *a grimasa* 'grimacer' (absent dans le roumain de nos jours) « il vaut mieux qu'à la place de ce verbe on emploie le nom accompagné par le verbe *facere* 'faire' » (s. v. *Grimassare*). La forme non accentuée de la première personne du pluriel du pronom personnel complément d'objet direct ou indirect du verbe, *ne* 'ne, nous', « se place après le verbe quand celui-ci est à l'impératif […], et devant le verbe dans tous les autres cas » (s. v. *Ne*). Enfin, en tant que « marqueur de négation », la particule *ne* 'non' « s'exprime en français à l'aide de *il-*, *im-*, *in-*, *ir-*, *mé-*, *non* [… et e]lle accompagne presque tous les supins des verbes qu'elle transforme dans ce cas en adjectifs, presque tous les infinitifs qu'elle transforme en noms et presque tous les adjectifs qualificatifs. Elle reste toujours devant eux et elle ne se sépare jamais du mot qu'elle accompagne » (s. v. *Ne*), etc.

2.1.3 Niveau lexical

Du point de vue lexical, nous constatons que les indications apportant des précisions sur la circulation d'un mot, sur son statut dans la langue (mot obsolète ou avec un emploi restreint) ou sur son sémantisme (sens propre ou figuré) sont les plus nombreuses dans les deux dictionnaires.

Du LBe, nous mentionnons, à titre d'exemple, l'emploi particulier de l'adjectif *primariu* 'initial, originaire; germain' qui est décrit dans une notice de la façon suivante: « *Nota.* Cet *adjectif* (comme d'autres mots) nous montre qu'au cours du temps, de nombreux mots d'origine latine ont perdu chez les Roumains leur sens originaire, en s'employant actuellement

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

seulement dans certaines constructions figées ou particulières, ainsi l'adjectif *primariu* 'germain (cousin ~)', je l'ai entendu s'employer uniquement dans ces deux situations » (s. v. *Primariu*). Le mot à circulation régionale *săbău* 'tailleur' est « un mot qui ne s'emploie qu'en Hongrie, *i. e.* tailleur » (s. v. *Sâbȏu*), tandis que *şineag* 'outil pour mesurer une certaine quantité de produits agricoles' est en circulation uniquement « dans le Banat » (s. v. *Şinégu*). Le mot *babă* est d'habitude employé pour désigner soit *une guérisseuse*, soit *une vieille femme*, mais les rédacteurs du LBe attirent l'attention sur le fait que « dans certaines parties du Pays hongrois on emploie ce mot à la place du mot *mère* » (s. v. *Babâ*). Pour le mot *merţă* 'boisseau' les rédacteurs notent que « dans certains endroits, il signifie double décalitre, […] tandis que dans d'autres endroits il signifie un demi-double décalitre » (s. v. *Merţâ*), etc.

On attire également l'attention du lecteur sur l'habitude des Roumains de créer des dénominations locales pour différents objets (au sens large) de la réalité à la place d'employer les noms déjà existants; c'est le cas de la confusion créée entre *petits pois* et *haricots*, confusion qui est présentée de la manière suivante: « *Nota.* Beaucoup de gens comprennent par pois les haricots verts et pour les distinguer quand même ils appellent les haricots 'pois de jardin' ou 'pois grimpants', tandis que les petits pois proprement dits, ils les appellent petits boutons ou perles-des-champs, mais cette confusion ne se justifie pas car il vaut toujours mieux de distinguer les espèces par leurs propres dénominations » (s. v. *Mazere*). Mentionnons aussi l'emploi erroné du mot latin *canistrum* pour désigner *straiţa* 'la besace': « *Nota.* la plupart des gens désignent *la besace* par le mot latin *canistrum*; chez Cicéron, Virgile ou Ovide, ce mot ne signifie pas besace, mais panier ou corbeille où l'on met le pain » (s. v. *Straiţâ*), etc.

Bien que le sens propre ou figuré soit signalé au sein des définitions mêmes à travers les mots latins *proprie* 'propre', respectivement *metaphorice* ou *tropice* 'métaphorique, figuré', comme dans les exemples qui suivent [le sens propre du syntagme *piciorul caprii* 'le pied, la jambe de la chèvre' est marqué par les rédacteurs ainsi « a) *proprie*: *pes caprinus* […] » (s. v. *Piciorul' caprii*), tandis que le sens figuré du mot *toiag* 'bâton' est souligné par la formule « *tropice*: le bâton de vieillesse, c'est-à-dire le support, l'appui […] » (s. v. *Toégu*), etc.], nous avons identifié aussi des structures phraséologiques introduites par la formule abrégée « NB. » ou « Not. »; par exemple, la construction *a cădea din cer* 'tomber du ciel', au-delà de son sens littéral, acquiert une sens figuré quand elle s'emploie en tant qu'épithète pour décrire une personne: « NB. Tombé du ciel se dit de quelqu'un qui est aimé et bon et qui arrive au bon moment » (s. v. *Câdzutu*).

En revanche, dans le VRF, les articles renfermant des indications sur l'usage lexical de tel ou tel mot visent surtout la circulation en parallèle de deux formes lexicales pour le même référent: le nom *trompă* 'trompe' « se dit parfois à la place du mot *trompetă* 'trompette' » (s. v. *Trompă*), tandis que le nom hétérogène *revirement* 'revirement' « [s]e dit surtout *virement* 'virement' » (s. v. *Revirementŭ*); bien que le référent soit identique, il y a parfois des cas où l'on change de forme selon le registre: le mot *coşciug* 'cercueil' « s'emploie pour n'importe quelle personne décédée », tandis que le mot *şicriu* 'cercueil, catafalque' « est à employer dans le cas des saints et, par extension, des personnalités » (s. v. *Şicriŭ*), etc. Dans certains cas, la définition d'un mot se voit enrichie de marques diatopiques ou temporelles, censées aider le lecteur ou l'apprenant à s'approprier correctement le sémantisme du mot: le nom *urnă* 'urne' bénéficie pour sa première définition d'une délimitation géographique: « [il s'agit,] chez les Romains, d'une unité de mesure pour le volume des liquides » (s. v. *Urnă*); le nom féminin *tribună* 'tribune' connaît au cours du temps des sens différents: « dans l'Antiquité, un emplacement élevé d'où les orateurs grecs et romains s'adressaient au peuple. – De nos jours, le mot désigne une galerie de bancs dans les salles de réunions publiques pour les

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

orateurs ou pour le public […]. – Dans les églises et dans certains endroits publics, une plateforme élevée où se trouvait l'orchestre » (s. v. *Tribună*), etc.

Pour d'autres mots, le rédacteur préfère signaler leur emploi restreint après la définition, Ainsi, le nom *toxic* 'toxique' « est employé seulement dans les cercles savants » (s. v. *Toxicŭ*), l'adjectif *rânzos* 'folliculaire', peu usité et employé au figuré, « se dit à propos des enfants et des hommes qui se fâchent facilement […] » (s. v. *Rînzossŭ-ă*), etc.

Pour une meilleure fixation du sémantisme du mot, le rédacteur recourt à des formules du type « este opusul lui… » 'c'est l'antonyme de…' ou « sinonim lui… » 'synonyme de …'. Par exemple, dans le cas du nom *teism* 'théisme', juste après la définition on affirme: « c'est l'antonyme d'*ateism* 'athéisme' » (s. v. *Teismŭ*), pour le nom *tartru* 'tartre', juste avant la définition, le rédacteur marque: « synonyme de *tartrat* 'tartrate' » (s. v. *Tartrŭ*), ou pour l'adjectif *presinte* 'présent', on consigne que « c'est l'antonyme d'*absinte* 'absent' » (s. v. *Presinte-ă*), etc.

3. En guise de conclusion

L'analyse de notre corpus nous permet d'en tirer quelques conclusions.

1) Comparant les notices d'ordre normatif présentes dans environ 400 articles puisés dans le LBe sur plus de 12.000 entrées (soit 3,33%) avec celles identifiées dans environ 1800 articles repris du VRF sur environ 28.000 entrées (soit 6,42%), nous constatons une évolution au niveau de la description du mot; il s'agit d'un passage d'un mode de description du type commentaire introduit parfois par les formules latines *nota* ou *usu* (le cas du LBe), vers un mode de description plus synthétique mis en évidence par des parenthèses rondes (le cas du VRF).

2) Le type d'informations contenues dans les notices normatives concerne tous les niveaux de la langue, i. e. phonétique, morphologique, syntaxique et lexical. Dans le cas du LBe les indications normatives les plus nombreuses visent les niveau lexical, suivi par le niveau morphologique et celui phonétique, tandis que dans le cas du VRF la première place est occupée en égale mesure par le niveau morphologique et celui lexical, suivis par celui phonétique.

3) Cette approche descriptive nous a permis d'avoir une nouvelle perspective d'une part sur l'évolution lexico-grammaticale et sémantique des mots, tout en apprenant des informations précieuses sur la norme, la variation linguistique et surtout la dynamique de la langue roumaine au XIXe siècle (telle qu'elle est consignée et fixée par ces deux dictionnaires au cours de presque un demi-siècle), et d'autre part sur la manière dont ces notices sont enregistrées et sur le niveau de la langue qu'elles touchent, en mettant en évidence aussi bien leur fonction didactique, que le rôle du dictionnaire dans l'apprentissage de la langue maternelle. Au-delà de l'organisation macro- et micro-structurelle du dictionnaire et des articles, ce fait souligne également l'attitude envers la langue que de tels rédacteurs d'ouvrages lexicographiques présentent à un moment donné de l'histoire, de même que la manière dont leur écriture rédactionnelle reflète aussi leurs goûts culturels.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Références

Aldea, M. (2018a): L'enjeu de l'orthographe dans le processus d'affirmation de la langue roumaine. In: Antonelli, R./Glessgen, M./Videsott, P. (ed.): Atti del XXVIII Congresso internazionale di linguistica e filologia romanza (Roma, 18-23 luglio 2016), volume 2. Strasbourg, p. 1255–1265.

Aldea, M. (2018b): Alfabetul limbii române: o abordare lexicografică. In: Lupu, C. (ed.): Ciolan, A./Zuliani, A. (coed.): Studii romanice. I. Omagiu profesorilor Florica Dimitrescu şi Alexandru Niculescu la 90 de ani. Bucureşti, p. 17–27.

Atkins, S. B. T./Rundell, M. (2008): The Oxford guide to practical lexicography. Oxford.

Cabré, Ma. T. (1998): La termionologie. Théorie, méthode et applications. Ottawa/Paris.

Chivu, Gh. (2019): Limbă şi cultură. Studii de istorie a limbii române literare. Bucureşti.

Costinescu, M. (1979): Normele limbii literare în gramaticile româneşti. Bucureşti.

Dubois, J. (1970): Dictionnaire et discours didactique. In: Langages 19,pp. 35–47.

Dubois, J./Dubois, Cl. (1971): Introduction à la lexicographie. Le dictionnaire. Paris.

Gafton, A. (2012): De la traducere la norma literară. Contribuţia traducerii textului biblic la constituirea vechii norme literare. Iaşi.

Gheţie, I. (1975): Baza dialectală a românei literare. Bucureşti.

Ghibu, O. (1998): Din istoria literaturii didactice româneşti. Bucureşti.

LBe (2013): https://doi.org/10.26424/lexiconuldelabuda (dernière consultation: 30-01-2022).

Svensén, Bo. (1993): Practical lexicography. Principles and methods of dictionary-making. Oxford/New York.

Seche, M. (1966): Schiță de istorie a lexicografiei române. București.

VRF (1870): I. Costinescu, Vocabularu romano-francesu. Vol. I–II. Bucuresci.

Zgusta, L. (1971): Manual of lexicography. The Hague/Paris.

Informations de contact

Maria Aldea
Université Babeş-Bolyai de Cluj-Napoca
maria.aldea@ubbcluj.ro

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Harald Bichlmeier

ALMANCA TUHFE / DEUTSCHES GESCHENK (1916) oder: Wie schreibt man deutsch mit arabischen Buchstaben?

Abstract Versified dictionaries are bilingual/multilingual glossaries written in verse form to teach essential words in any foreign language. In Islamic culture, versified dictionaries were produced to teach the Arabic language to the young generations of Muslim communities not native in Arabic. In the course of time, many bilingual/multilingual versified dictionaries were written in different languages throughout the Islamic world.

The focus of this study is on the Turkish-German versified dictionary titled *Almanca Tuhfe / Deutsches Geschenk* [German Gift], published by Dr. Sherefeddin Pasha in Istanbul in 1916. This dictionary is the only dictionary in verse ever written combining these two languages. Moreover the dictionary is one of the few texts containing German words written in Arabic letters (applying Ottoman spelling conventions). The study concentrates on the way German words are spelled and tries to find out, whether Sherefeddin Pasha applied something like fixed rules to write the German lexemes.

Keywords Osmanische Wörterbücher; Osmanisch-deutsche Wörterbücher; Wörterbücher in Versen; Ajamiado-Literatur; arabische Schrift; osmanisch-türkische Orthographie

1. Einleitung

Im Jahre 1916 veröffentlichte Dr. Scherefeddin Pascha in Istanbul/Constantinopel sein Osmanisch-Deutsches Wörterbuch *Almanca tuhfe / Deutsches Geschenk*. Bei diesem Wörterbuch handelt es sich um das einzige (osmanisch-)türkisch-deutsche Wörterbuch in Versen, das jemals im Druck erschienen ist. Es steht in der langen Tradition der in Versen verfassten Wörterbücher in der islamischen Welt, die anfangs in erster Linie dazu dienten, Sprechern nichtarabischer Sprachen das Arabische als Sprache des Korans nahezubringen. Ab dem 17. Jahrhundert wird diese Tradition von den Osmanen übernommen und ausgebaut und dient nun zur Vermittlung des Osmanisch-Türkischen an die nicht türkischen Völkerschaften im Osmanischen Reich, wie etwa Armenier, Griechen, Bosnier, Bulgaren, Albaner, und umgekehrt auch der Sprachen jener Völker an die Türken.[1]

Das erste bekannte derartige zweisprachige Glossar in Versen ist *Nisab al-Sibyan* (1221), das dazu gedacht war, arabische Wörter Sprechern des Persischen (Fārsī) zu vermitteln. Es wurde von Farahi († 1242) in Afghanistan verfasst. In der islamischen Tradition wurden ab dem 13. Jahrhundert bis ins 21. Jahrhundert derartige Glossare für zahlreiche Sprachen verfasst, so für Arabisch, Persisch, Türkisch, Hindi, Urdu, Kurdisch, Sanskrit, Pashto, Tabari, Bulgarisch, Griechisch, Armenisch, Bosnisch und Albanisch, ab dem 19. Jahrhundert entstanden auch zweisprachige Wörterbücher für Sprachen außerhalb des Osmanischen Reichs, so für Englisch, Französisch und eben Deutsch.[2]

[1] Vgl. ausführlich zu diesem Werk und seiner Stellung innerhalb der Tradition dieser Art Wörterbücher Doğan Averbek/Bichlmeier (2020, 2021), Bichlmeier/Doğan Averbek (2021).

[2] Vgl. Doğan Averbek (2018a, S. 87). Eine Auswahl aus den zahlreichen Werken sei im Folgenden angeführt: Türkisch-Arabisch: Abdüllatîf b. Melek, *Lugat-i Ferişteoğlu* (Text in Muhtar 1993), Sün-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Besonderheit dieser im Osmanischen Reich entstandenen Wörterbücher ist neben dem Umstand, dass sie eben in Versen abgefasst sind, um so mnemotechnisch den Erwerb der fremden Sprache zu unterstützen, dass die Wörter der jeweiligen nicht türkischen Sprache in arabischer Schrift und dann weitgehend nach den Regeln der osmanischen Orthographie verschriftet wurden.

Exkurs zur *Aljamiado*-Literatur

Was wir hier sehen – bei den deutschen Wörtern des Wörterbuchs – ist letztlich ein Ausläufer der sog. *Aljamiado*-Literatur, d. h. von Literatur (im weiteren Sinn des Wortes), die in arabischer Schrift geschrieben ist, in einer Sprache, für die die arabische Schrift gewöhnlich nicht verwendet wird. Der allgemein akzeptierte *Terminus technicus* beruht auf dem spanischen Wort *aljamiado*, einer Ableitung von spanisch *aljamía*, einem Lehnwort aus arabisch *al-ʾaʿǧamiya* ‚nicht arabisch, fremd'.[3]

Der Gebrauch der arabischen Schrift für Sprachen, die gewöhnlich in lateinischer, kyrillischer oder auch griechischer Schrift verschriftet wurden und werden, war in einer bestimmten Periode recht weit verbreitet. Sie wurde meist von islamisierten bzw. zum Islam übergetretenen Bevölkerungsgruppen verwendet, die unter arabische bzw. später osmanische Herrschaft gekommen waren: so z. B. von Sprechern des Spanischen im Mittelalter, von Sprechern von Varietäten des serbokroatischen Dialektkontinuums in Bosnien ab dem 16. Jahrhundert bis ins 20. Jahrhundert und Sprechern des Weißrussischen/Belarussischen ab dem 17. Jahrhundert bis in die Mitte des 20. Jahrhunderts. Der Fall dieser weißrussischen Texte ist noch einmal etwas anders gelagert als der der Spanier und Bosnier: Hier haben islamisierte turksprachige Tataren, die im Spätmittelalter als Söldner aus der Region der Krim in den Norden von Polen-Litauen gebracht worden waren, den Sprachwechsel zum Weißrussischen vollzogen.[4]

Bosnische Texte in arabischer Schrift wurden ab dem Ende des 19. Jahrhunderts in Bosnien-Herzegowina, das seit 1878 von Österreich-Ungarn okkupiert war und 1908 auch annektiert wurde, auch gedruckt. Österreich-Ungarn unterstützte damit die Entstehung einer bosnisch-muslimischen Identität, um ein Gegengewicht gegen den wachsenden kroatischen und serbischen Nationalismus in der Region zu schaffen.[5] Die Drucke wurden auch nach 1918 im neugegründeten Jugoslawien fortgesetzt, endeten aber mit der deutschen Invasion

bülzâde Vehbî, *Nuhbe-i Vehbî* (Text in Yurtseven 2003), Mehmed b. Ahmed er-Rumî (?), *Sübha-i Sıbyân* (Text in Kılıç 2007, S. 29–71), Fedâyî Mehmed Dede, *Tuhfe-i Fedâyî* (Text in Fedayî 2019); Türkisch-Persisch: Hasan Rızâyî, *Kân-ı Meʿânî* (Text in Turan 2012, S. 2939–2992), Hüsâm b. Hasan el-Konevî, *Tuhfe-i Hüsâmî ez Mültekât-ı Sâmî* (Text in Arslan 2016), Şâhidî İbrâhîm Dede, *Tuhfe-i Şâhidî* (Text in Verburg 1997, S. 5–87), Sünbülzâde Vehbî, *Tuhfe-i Vehbî* (Text in Vehbî 2012); Türkisch-Arabisch-Persisch: Hâkî Mustafâ-yı Üsküdarî, *Menâzımü'l-Cevâhir* (Text in Arslan 2011), Osman Şâkir b. Mustafâ Bozokî (Şâkirî), *Müselles-nâme-i Şâkir* (Text in Kaya/Ayçiçeği 2019); Türkisch-Albanisch: Nazîm, *Der Beyân-ı Türkî maʿa Lisân-ı Arnabud* (Text in Rossi 1946, S. 219–246), Mahmûd, *Dürre-i Manzûme* (Untersuchung in Doğan Averbek 2018b); Türkisch-Bosnisch: Muhammed Hevâ'î Üsküfî, *Makbûl-ı Ârif* (Text in Uskufi 2001); Türkisch-Armenisch: Ref'î Kâlâyî, *Manzûm Lugat-ı Ermeniyye* (Text in Dankoff/Kut/Weitenberg 1996); Türkisch-Griechisch: Hanyalı Osmân Nûrî, *Lugat-ı Manzûme-i Nûriyye berây-ı Terceme-i Lisân-ı Rumiyye* (Text in Erik 1982); Türkisch-Französisch: Yûsuf Hâlis Efendi, *Miftâh-ı Lisân* (Text in Yûsuf Hâlis 2006).

3 Vgl. Kontzi (1980); Hadžijahić (1941).

4 Zur Geschichte der Tataren in Polen-Litauen vgl. Bairašauskaitė (2021).

5 Vgl. Neweklowsky (1996, S. 63–67).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Jugoslawiens im April 1941. Versuche, die Tradition des Schreibens in der sog. *Arebica*[6] nach dem Zerfall Jugoslawiens (1991–1992) und dem Bürgerkrieg (1992–1995) haben bislang wenige greifbare Ergebnisse gebracht. Selbst innerhalb der islamischen Gemeinschaft in Bosnien ist das Konzept, Bosnisch in arabischer Schrift zu schreiben nicht mehrheitsfähig.

Weißrussische Texte sind ab dem 16. Jahrhundert handschriftlich bezeugt,[7] in arabischer Schrift gedruckt wurden solche Texte vor dem Zweiten Weltkrieg in Polen, doch beendete der Zweite Weltkrieg diese Tradition ebenso wie in Bosnien. Der Stadtrat von Warschau hatte im Frühjahr 1939 noch den Bau einer Moschee für die Tataren genehmigt, woraus ebenfalls nichts mehr wurde.

2. Zu *Almanca tuhfe / Deutsches Geschenk*

Das Büchlein *Almanca tuhfe / Deutsches Geschenk* wurde 1916 in Istanbul/Constantinopel gedruckt und kann damit als Ergebnis der sich seit Ende des 19. Jahrhunderts immer intensiver entwickelnden Deutsch-Osmanischen Militär- und Kulturkontakte und schließlich auch Bündnisse gesehen werden. Diese Verbindungen befanden sich während des Ersten Weltkriegs auf einem Höhepunkt.[8] Das vorliegende Glossar kann sicher als Ausfluss der Waffenbrüderschaft gewertet werden. Ziel des Büchleins war es, deutschen Elementarwortschatz zu vermitteln.

Die Druckauflage des Buchs ist nicht bekannt. Es ist heute jedenfalls eine Rarität. Bislang konnten in Bibliotheken der Türkei (Erzurum) und Deutschlands (München) je ein Exemplar verifiziert werden, eines fand sich im antiquarischen Buchhandel.

Über den Autor Scherefeddin Pasha (Şerefeddin Paşa) ist wenig bekannt: Er war Militärarzt in der Osmanischen Armee und diente u. a. 1908 im Militärhospital in Zeytinburnu[9] und vor 1916 als Direktor des Militärkrankenhauses in Van. In einer Zeitungsmeldung von 1926 wird sein Ableben vemeldet.[10] Der Autor schreibt im Vorwort seines Buchs, dass er in Privatstunden Deutsch gelernt habe, weil er die Notwendigkeit erkannt habe, dass ein solches Werk verfasst werden müsse.[11]

3. Struktur und Inhalt von *Almanca tuhfe*

Das Buch enthält auf 52 Seiten knapp 2.000 (ca. 1.200 verschiedene) deutsche Wörter in etwa 1.300 Einträgen von Einzelwörtern, Wortfügungen oder Kurzsätzen.

Das Buch beginnt mit einer Prosa-Präambel und besteht aus 23 Abschnitten, die sich aus einem Teil *sabab-i nazm* (Grund für die Abfassung des Gedichts), 21 Abschnitten Glossar-

6 Vgl. Hadžijahić (1955).

7 Vgl. als Ausgabe eines handschriftlichen Textes aus der ersten Hälfte des 18. Jahrhunderts Miškinienė et al. (2009).

8 Vgl. aus der überbordenden Literatur zum Thema exemplarisch Grüßhaber (2018).

9 Vgl. Yıldırım (2006, S. 347). Seine Arbeit dort ist auch archivalisch belegt (Staatsarchive der Republik Türkei, 253–257, Y.PRK.ASK).

10 Vgl. Haber (1926, S. 4).

11 Scherefeddin Pascha (1916, S. 3 f.).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Teil und einem Epilog aus zwei Strophen zusammensetzen. Das Werk besteht insgesamt aus drei einzelnenen Strophen und 295 Couplets.[12]

Die einzelnen Abschnitte tragen die folgenden Titel und enthalten die angegebene Zahl von Strophen:

	türkischer Titel	**deutscher Titel (in Schreibweise des Originals)**	**Zahl der Couplets**
1.	*Kâ'inât*	*Die Wesen*	11
2.	*E'âzım*	*Die Grossen*	7
3.	*Ebeveyn hısım ve akrabâ*	*Die Eltern*	7
4.	*Sayılar*	*Die Zahlen*	17
5.	*Hafta Eyyâmı*	*Die Wochen Tage*	3
6.	*Senenin on iki ayları*	*Die Monate Des Jahres*	4
7.	*Sıfât-ı muhtelife*	*Das Bei Wort*	17
8.	*Me'kûlât ve meşrûbât*	*Die Speise Und Getränke* [sic]	15
9.	*Yemişler ve sebzeler*	*Die Früchte Und Gemüse*	13
10.	*Hayvânât – kuşlar – balıklar*	*Die Thiere Und Vögel Und Fische*	15
11.	*A'zâ-yı beden*	*der Menschliche Körper* [sic]	13
12.	*Melbûsât*	*Die Kleidungen*	10
13.	*Mesken – ev eşyâsı*	*Wohnung-Möbel*	18
14.	*Almanca fi'ller*	*Das Zeit Wort*	51
15.	*Mekteb*	*Die Schule*	15
16.	*Lugât-i muhtelife*	–	18
17.	*Harf-i cerler*	*Das Vorwort*	12
18.	*Zamîrler*	*Die Pronomen*	15
19.	*Meşhûr şehrler*	*Die Städte*	6
20.	*Mükâleme*	*Die Spräche*	24
21.	*Sâ'atlere dâ'ir*	*Von der Uhr*	4

Tab. 1: Struktur von *Almanca tuhfe*

Das Wörterbuch Schereffeddin Paschas zeigt die auch sonst bekannten Schwierigkeiten bei der erstmaligen Verschriftung von Sprachen mit einem neuen Schriftsystem. Zwar gab es frühere Verschriftungsversuche des Deutschen mit arabischer Schrift (etwa eine Sammelhandschrift vom Ende des 16. Jahrhunderts, die neben deutschen Texten auch lateinische, ungarische, kroatische Texte in arabischer Schrift enthielt und die die phonetischen/phonologischen Gegebenheiten der transkribierten Sprachen genauer wiederzugeben versuchte, als dies hier geschehen ist),[13] doch hat sich weder aus jenen älteren Versuchen noch aus dem

[12] Mittlerweile liegt eine Publikation des Wörterbuchs vor (Doğan Averbek/Bichlmeier 2020), die das Faksimile und die Transliteration des gesamten Textes sowie einen Kurzkommentar zu den deutschen Wörtern samt Analyse der ‚Schreibregeln' und eine Einführung zu versifizierten Wörterbüchern als Lehrbücher enthält.

[13] Zum Kulturellen und ethnischen Hintergrund des Verfassers dieser Handschrift vgl. Römer (2014). Zur Handschrift und zu Schreibgewohnheiten darin vgl. Mittwoch/Mordtmann (1927); (aus ungarologischer

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

jüngeren von 1916 eine Tradition der Verschriftung des Deutschen in arabischer Schrift herausgebildet, die brauchbar und schließlich auch massentauglich gewesen wäre. Mit dem Zusammenbruch des Osmanischen Reichs im Gefolge des verlorenen Ersten Weltkriegs und dem wenige Jahre später erfolgten Übergang der Türkei von der arabischen Schrift zur Lateinschrift wurden weitere derartige Versuche ohnehin überflüssig.

Bei der Lektüre wird schnell deutlich, dass Scherefeddin Pascha kein völlig durchdachtes System zur Verschriftung des Deutschen mit arabischen Buchstaben in osmanischer Orthographie verwendet hat:

Einerseits orientiert er sich am deutschen Schriftbild, wenn er etwa dt. <ck> inlautend mit arab. <qk> wiedergibt, andererseits jedoch an der deutschen Phonetik bzw. Phonologie, wenn (silben-)auslautendes dt. <ck> dann als arab. <k> verschriftet wird oder deutsche geschriebene Geminaten in der Regel mit einfachen arabischen Buchstaben wiedergegeben werden. Auch bei den Vokalen ist die Darstellung nicht systematisch, so wird etwa der dem Türkischen fremde, im Deutschen aber phonologisch relevante Gegensatz Langvokal vs. Kurzvokal ignoriert. Und selbst die im Prinzip im arabischen Alphabet mögliche genaue Vokalmarkierung mittels diakritischer Zeichen wurde ebensowenig zur Anwendung gebracht wie die in der osmanischen Orthographie übliche konsequente Verwendung von den Graphemen für einfache Konsonanten vor vorderen Vokalen gegenüber der Verwendung von Graphemen für sogenannte ‚emphatische' Konsonanten vor hinteren Vokalen.

4. Das Transliterationssystem der deutschen Wörter

In den deutschen Wörtern wird der Akzent nicht markiert.

Bei den Vokalen ist die dort existierende Längenkorrelation nicht in der Transkription reflektiert. Scherefeddin Pascha war sich möglicherweise dieser Option gar nicht bewusst, da sie im Türkischen nicht existiert.

Insgesamt ergibt sich folgendes Bild:

4.1 Arabisch-deutsche Entsprechungen

Arabisches Graphem	Transliteration	phonemische Entsprechung im Deutschen
ب	**b**	**/b/:** <bet> *Bett*, <zun ābnd> *Sonnabend*
پ	**p**	**/p/:** <pfrd> *Pferd*, <šupfer> *Schöpfer*, <šperber> *Sperber*
ت	**t**	**/t/:** <tuḫtr> *Tochter*, <ārtišuqke> *Artischocke*, <mitwuḫ> *Mittwoch*
ث	**ṯ**	–
ج	**ǧ**	–

Sicht) Gragger (1927a, 1927b); Ivušić (2012); Bichlmeier/Ivušić (2013); Ivušić (2013); Bichlmeier (2020). Zu Editionen des Manuskripts siehe Blau (1868); Mittwoch (1927). Das Manuskript ist in orientalistischer Terimologie als *majmua* (türk. *mecmua*), also als Sammelhandschrift gemischten Inhalts zu bezeichnen.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Arabisches Graphem	Trans-literation	phonemische Entsprechung im Deutschen
چ	**č**	**/ʧ/ <tsch>** (nur an Morphemgrenzen und in Lehnwörtern): <bučafter> *Botschafter*, <dulmečr> *Dolmetscher* **/ts/ <z>:** <čen> *zehn*, <čan šmrč> *Zahnschmerz*, <čan> *Zahn*, <čunge> *Zunge*, <hrč weḥ> *Herzweh* **/s/ in /ts/:** nur in <qatče> *Katze*
ح	**ḥ**	für ‚stilles' <h> /-Ø/ im Auslaut: <quḥ> *Kuh*, <šuḥ> *Schuh*, <fluḥ> *Floh* [flo:], für Auslautendes **/x/ <ch> [ç]:** in <milḥ> *Milch* und Komposita mit diesem Wort im Namen <Muḥammed> *Mohammed*
خ	**ḫ**	**/x/ <ch> [χ] / [x] / [ç]** (im Inlaut und Auslaut): <ziḫ> *sich*, <mitwuḫ> *Mittwoch*, <huḫ> *hoch*, <naḫ> *nach*, <quḫin> *Kuchen*, <maḫin> *machen* Manchmal (als Druckfehler?) für ‚stilles' <h>: in <qupf weḫ> *Kopfweh* (vs. <hrč weḥ> *Herzweh*), <fruḫ> *früh*
د	**d**	**/d/:** <diq darmy> *Dickdarm*, <zun ābnd> *Sonnabend*
ذ	**ḏ**	–
ر	**r**	**/r/:** <fruḫ> *früh*, <hrč weḥ> *Herzweh*, <šupfer> *Schöpfer*, <šperber> *Sperber*
ز	**z**	**/s/ [z]** (im Anlaut/Silbenonset und zwischen Vokalen): <zun ābnd> *Sonnabend*, <zu> *so*, <aunzr> *unser*, <haze> *Hase* **/z/ [s]** im Auslaut, wenn es einen Wechsel [s] > [z] im Paradigma gibt (vgl. Dat. <hauze> *Hause*): <hauz> *Haus*, <airrin hauz> *Irrenhaus*, <qranqn hauz> *Krankenhaus*, einziges Bsp., das nicht *Haus* enthält: Pers.Pron. <ᶜunz> *uns* (vielleicht nach dem Modell von <aunzr> *unser* ?)
ژ	**ž**	–
س	**s**	**/s/ [s]** im Auslaut und manchmal im Inlaut, ausnahmsweise im Anlaut: <salat> *Salat*,[14] <zamistaġ> *Samstag*
ش	**š**	**/ʃ/ <sch>,** auch für <s> in <sp, st> im Silbenonset: <ārtišuqke> *Artischocke*, <šupfer> *Schöpfer*, <šperber> *Sperber*
ص	**ṣ**	–
ض	**ḍ**	–
ط	**ṭ**	**/t/:** <ġranaṭe> *Granate* (‚Granatapfel'), manchmal **/d/** im Inlaut wie in <limunaṭe> *Limonade*; wahrscheinlich als Übernahme aus der arabischen/osmanischen Schreibung in <slṭan> *Sültan*
ظ	**ẓ**	–

[14] Vgl. ebenso <sardine> *Sardine*, <sufa> *Sofa*, <suldat> *Soldat*; im Standarddeutschen gibt es [s] im Anlaut vor Vokal nicht. Interessanterweise sind all diese Wörter Lehnwörter. Nach Siebs (1905, S. 60; 1915, S. 66 f.) soll stimmloses antivokalisches [s-] in Wörtern französischer oder italienischer Herkunft gesprochen werden, wenn sie noch als Fremdwörter gelten, aber stimmhaft, wenn sie schon als eingedeutscht eingestuft werden. In Wörtern griechischen und lateinischen Ursprungs ist [z-] die dominante Aussprache. Unter den letztgenannten Wörtern listet Siebs explizit *Salat*; folglich sollte die Aussprache [zaˈla:tʰ] (gewesen) sein, während Scherefeddin Pascha offenbar an eine Aussprache /sala:t/ [saˈla:tʰ] denkt.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Arabisches Graphem	Trans-literation	phonemische Entsprechung im Deutschen
ع	**c**	nur in bestimmten Kombinationen, s.u.
غ	**ġ**	**/g/ [g], [ç]/i_** etc., bes. im Auslaut: <wiḫtiġ> *wichtig*, <zalčiġ> *salzig* (oder Schreibung nach der dt. Vorlage?) oft nach /a/: <fraytaġ> *Freitag*, <zamistaġ> *Samstag*
ف	**f**	**/f/:** <fraytaġ> *Freitag*
ق	**q**	**/k/** im Anlaut: <qauen> *kauen*, <qalb> *Kalb*, <qamel> *Kameel* (mod. nhd. *Kamel*) als erster Teil des Digraphen <qk>, der dt. <ck> wiedergibt (jedoch nicht durchgängig: <diq darmy> *Dickdarm* [zur Vermeidung eines Konsonantenclusters aus drei Graphemen?])
گ	**g**	**/g/** (im Anlaut und Inlaut): <general> *General*, <gezund> *gesund*, <gelb> *gelb*, <flige> *Fliege*, <faige> *Feige* selten **/g/** im Auslaut: <fus weg> *Fußweg*
ل	**l**	**/l/:** <lam> *Lamm*
م	**m**	**/m/:** <lam> *Lamm*, <mutter> *Mutter*
ن	**n**	**/n/:** <gezund> *gesund*
ه	**h** **e**	**/h/:** <huze> *Hose*, <huḫ> *hoch* **/e/ [ɛ]** (im Inlaut): <šperber> *Sperber*, <fet> *fett* **/e/ [ə]** (in unbetonten Silben, bes. im Auslaut und in Präfixen, Präverbien): <qauen> *kauen*, <tauen> *tauen*, <sardine> *Sardine*, <gezund> *gesund* **/ē/ [e:]** (im Inlaut): <qamel> *Kameel* (mod.nhd. *Kamel*), <leber> *Leber*, <kele> *Kehle* **/ē/ [e:]** (im Auslaut): <fe> **/ǟ/ [ɛ:]:** <šedel> *Schädel*
ي	**i**	**/i/:** <diq darmy> *Dickdarm* **/ī/:** <zibin> *sieben* **/j/** (als zweiter Teil in Diphthongen): <zain> *sein* **/j/** (im Anlaut vor Vokalen): <iares> *Jahres*, <iaqke> *Jacke*, <iuny> *Juni*, <iujy> *Juli*, <ianuar> *Januar*, <ist> [Verschreibung für *<ičt>] *jetzt*, <iunġ> *jung* **/ə/** (in unbetonten Silben): <āuġin brauuᵊe> *Augenbraue*, <maġin> *Magen*, <zibin> *sieben*, gewöhnlich in der Endung des Infinitivs *-en*/-ən/ [-ən], [-n̩]
ى	**y**	**/i/** (im Inlaut, selten: es repräsentiert unbetontes [i] in Fremdwörtern): <mynister> *Minister*, <auffyčir> *Offizier* **/ī/** (im Inlaut [selten] und im Auslaut von Einsilblern): <lybin> *lieben*, <wy> *wie*, <zy> *sie*, <dy> *die* **/j/** (als zweiter Teil von Diphthongen [selten]): <šraybin> *schreiben*, <ārbaytin> *arbeiten*

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Arabisches Graphem	Trans-literation	phonemische Entsprechung im Deutschen
و	**u**	**/u/** (im Inlaut): <mutter> *Mutter*, <hund> *Hund* **/ū/** (im Inlaut): <hun> *Huhn* **/o/** (im Inlaut): <tuḫtr> *Tochter*, <ārtišuqke> *Artischocke* **/ō/** (im Inlaut): <zun> *Sohn*, <huze> *Hose*, <huḫ> *hoch*, <ᵓelbuġin> *Ellbogen* **/ü/** (im Inlaut): <muqke> *Mücke* **/ü/** (im Anlaut): <ubersessin> *übersetzen* **/ǖ/** (im Inlaut): <zus> *süß*, <tur> *Tür* **/ö/** (im Inlaut): <šupfer> *Schöpfer* **/ȫ/** (im Inlaut): <fugel> *Vögel*, <šun> *schön*
	w[15]	**/v/** (im Anlaut): <wasser> *Wasser*
ا	**a**	**/a/** (im Inlaut): <wasser> *Wasser*, <qalb> *Kalb* **/a/** (im Auslaut): <sufa> *Sofa* **/ā/** (im Inlaut): <han> *Hahn*, <haze> *Hase*, <suldat> *Soldat* **/e/ [ʔɛ]** (im Anlaut): <anteriḫ> *Enterich*, <as> *es*, <altrny> *Eltern*
آ	**ā**	**/a/ [ʔa]** (im Anlaut oder im Silbenonset): <āffe> *Affe*, <ārbaytin> *arbeiten*, <ālt> *alt*, <ᶜum ārmin> *umarmen* **/ā/ [ʔaː]** (im Anlaut oder im Silbenonset): <ābnd> *Abend*, <zun ābnd> *Sonnabend*, <lineāl> *Lineal*
أ	**ᵓa**	**/e/ [ʔɛ]** (im Anlaut) <ᵓarbarmndn> *Erbarmenden* (Gen.)
او	**au**	**/o/ [ʔɔ]** (im Anlaut): <aunqel> *Onkel*, <auffyčir> *Offizier* **/ō/ [ʔoː]** (im Anlaut): <aur> *Ohr*, <aubr ārč> *Oberarzt* **/au/** (im Inlaut/Auslaut): <pfau> *Pfau*, <qranqn hauz> *Krankenhaus* **/ǖ/ [ʔʏː]** (im Anlaut): <aubrmurġn> *übermorgen* **/u/ [ʔʊ]** (im Anlaut): <aunzr> *unser*
اوي	**aui**	**/oi/ [ʔɔɪ]** (im Anlaut): <auier> *euer*, <auirer> *eurer*, <auiḫ> *euch*
اوو	**auu**	**/au/** (im Inlaut und Auslaut): <blauu> *blau*, <bauuḫ> *Bauch*
آوو	**āuu**	**/au/ [ʔaʊ]** (im [akzentuierten] Anlaut): <āuuster> *Auster*, <āuusgehin> *ausgehen*
آو	**āu**	**/au/ [ʔaʊ]** (im [unakzentuierten?] Anlaut): <āuġust> *August* aber <āuge> *Auge*, <āuġin ārč> *Augenarzt*
اي	**ai**	**/i/ [ʔɪ]** (im Anlaut): <aist> *ist*, <airrin hauz> *Irrenhaus* **/ī/ [ʔiː]** (im Anlaut): <ainen>, <ainn> *ihnen*
اي	**ai**	**/ai/** (im Inlaut): <laiḫt> *leicht*, <qraide> *Kreide*, <šun šray bin> *schönschreiben*
اى	**ay**	**/ai/** (im Auslaut): ārčnay> *Arznei*, <dray> *drei*, <bay> *bei*, <zay> *sei* **/ai/** (im Inlaut, falls Silbenauslaut): <blay feder> *Bleifeder*, <blay štift> *Bleistift*
اى	**ay**	**/ai/** (im Inlaut [selten]): <drayčen> *dreizehn* (Schreibung wie beim Simplex <dray> *drei*?)
ايي	**aii**	**/ai/** (im Inlaut): <drayis siġ> *dreißig*
آي	**āi**	**/ai/ [ʔaɪ]** (im Anlaut): <āis> *Eis*

[15] Als Abweichung von dem Prinzip, dass immer ein arabisches Graphem mit genau einem lateinischen Graphem transkribiert wird, wird hier <w> für *wāw* in den Fällen verwendet, wenn dt. /v/ transkribiert wird. Ziel ist es, die Transkription lesbarer zu machen.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Arabisches Graphem	Trans-literation	phonemische Entsprechung im Deutschen
آی	**āy**	**/ai/** (im Auslaut). <drāy> *drei*
عو	**ʿu**	**/u/ [ʔʊ]** (im Anlaut): <ʿund> *und*, <ʿunterhuze> *Unterhose*, <ziḫ ʿuntrhaltin> *sich unterhalten*, <ʿum> *um*, <ʿum ārmin> *umarmen* **/u/ [ʔuː]** (im Anlaut): <ʿur ġrus fater> *Urgroßvater*, <ʿur ġrus mutter> *Urgroßmutter*, <ʿur> *Uhr* **/ū/ [ʔʏː]** (im Anlaut): <ʿuberruq> *Überrock* **/ō/ [ʔoː]** (im Anlaut): <ʿurnšmrč> *Ohrenschmerz*
ئو	**ʾu**	**/ȫ/ [ʔøː]** (im Silbenonset): <aulifin ʾul> *Olivenöl* **/ö/ [ʔœ]** (im Anlaut): <ʾufnin> *öffnen* **/ü̈/ [ʔʏː]** (im Anlaut): <ʾubunġ> *Übung*
ئه	**ʾe**	**/ɛ/ [ʔɛ]** (im Anlaut und im Silbenonset): <ʾelbuġin> *Ellbogen*, <ʾessn> *essen*, <fir ʾeqkiġ> *viereckig*
ئ	ʾ	**/ɛ/ [ʔɛ]** (im Anlaut, bes. vor *-r-* [im Verbpräfix *er-*]): <ʾrwartin> *erwarten*, <ʾrwaḫin> *erwachen*, <ʾrlrnin> *erlernen*, <ʾrmudin> *ermüden*, <ʾrinnrny> *erinnern*, <ʾs> *es* **/e/ [ʔɛ]** (im Silbenonset): <beʾndiġin> *beendigen*
وی	**uy**	**/oi/** <eu>: <tuyfel> *Teufel* **/oi/** <äu>: <bruytiġām> *Bräutigam*
وي	**ui**	**/oi/** <eu>: <duič> *deutsch*, <huite> *heute*
ویو	**uiu**	**/oi/** <eu>: <nuiun ʿund nuiun čiġ> *neunundneunzig*

4.2 Deutsch-arabische Entsprechungen

Deutsche Phoneme (mit ihren Allophonen)	Trans-litera-tion	Arabische graphemati-sche Entspre-chung(en)
/b/: <bet> *Bett*, <zun ābnd> *Sonnabend*	**b**	ب
/d/: <diq darmy> *Dickdarm*, <ʾrmudin> *ermüden*, <dray> *drei*, <zun ābnd> *Sonnabend*	**d**	د
/f/: <fraytaġ> *Freitag*, <fugel> *Vögel*	**f**	ف
/g/ [kh], [ç]/i_ etc.,[16] bes. im Auslaut: <wiḫtiġ> *wichtig*, <zalčiġ> *salzig* oft nach /a/: <fraytaġ> *Freitag*, <zamistaġ> *Samstag* **/g/** (im Anlaut und Inlaut): <general> *General*, <gezund> *gesund*, <gelb> *gelb*, <flige> *Fliege*, <faige> *Feige* selten **/g/** im Auslaut: <fus weg> *Fußweg*	**ġ** **g**	غ گ
/p/: <pfrd> *Pferd*, <šupfer> *Schöpfer*, <šperber> *Sperber*	**p**	پ
/t/: <tuḫtr> *Tochter*, <ārtišuqke> *Artischocke*, <mitwuḫ> *Mittwoch* **/t/:** <ġranaṭe> *Granate* (‚Granatapfel'), wahrscheinlich als Übernahme der arabischen/osmanischen Schreibung in <slṭan> *Sultan* (S. 7), *Sültan* (S. 54)	**t** **ṭ**	ت ط

[16] Die Endung *-ig* wird im Standarddeutschen und nördlichen Varietäten als [-ɪç] gesprochen, in südlichen Varietäten indes als [-ɪkʰ].

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Deutsche Phoneme (mit ihren Allophonen)	**Trans-litera-tion**	**Arabische graphemati-sche Entspre-chung(en)**
/h/: <huze> *Hose*, <huḫ> *hoch* im Namen <Muḥammed> *Mohammed*, wo die Schreibung mit <-mm-> aus dem Deutschen übernommen ist	**h** **ḥ**	ه ح
/-ø/ ‚stilles' <-h>: <quḫ> *Kuh*, <šuḫ> *Schuh*, <fluḫ> *Floh* manchmal (Schreibfehler?) in deutschen Wörtern: <qupf weḥ> *Kopfweh*, <fruḫ> *früh*	**ḥ** **ḫ**	ح خ
/ʧ/ <tsch> (nur an Morphemgrenzen und Lehnwörtern): <bučafter> *Botschafter*, <dulmečr> *Dolmetscher*	**č**	چ
/ts/ <z>: <čen> *zehn*, <čan šmrč> *Zahnschmerz*, <čan> *Zahn*, <čunge> *Zunge*, <hrč weḥ> *Herzweh* **/ts/ <tz>:** <zitsen> *sitzen* **/ts/ <tz>:** <qatče> *Katze*	**č** **ts** **tč**	چ تس تچ
/x/ <ch> [χ] / [x] / [ç] (im Inlaut und Auslaut): <tuḫtr> *Tochter*, <ziḫ> *sich*, <mitwuḫ> *Mittwoch*, <huḫ> *hoch*, <naḫ> *nach*, <quḫin> *Kuchen*, <maḫin> *machen*	**ḫ**	خ
/z/ [z] (im Anlaut/Silbenonset vor Vokalen und zwischen Vokalen): <zun ābnd> *Sonnabend*, <zu> *so*, <haze> *Hase*, <aunzr> *unser*	**z**	ز
/r/: <hrč weḥ> *Herzweh*, <bučafter> *Botschafter*, <dulmečr> *Dolmetscher*	**r**	ر
/s/ [s] im Auslaut und manchmal im Inlaut, ausnahmsweise im Anlaut: <as> *es*, <salat> *Salat*, <sardine> *Sardine*[17]	**s**	س
/ʃ/ <sch>, auch für <s> in <sp, st> im Silbenonset: <šperber> *Sperber*, <šedel> *Schädel*	**š**	ش
/k/ im Anlaut: <qalb> *Kalb*, <qauen> *kauen*, <qamel> *Kameel* (mod. nhd. *Kamel*) als erster Teil des Digraphen <qk> als Wiedergabe von <ck> (aber nicht durchgängig: <diq darmy> *Dickdarm*)	**q**	ق
/l/: <lam> *Lamm*, <gelb> *gelb*, <flige> *Fliege*, <qamel> *Kameel* (mod. nhd. *Kamel*), <leber> *Leber* , <kele> *Kehle*	**l**	ل
/m/: <lam> *Lamm*, <mutter> *Mutter*	**m**	م
/n/: <gezund> *gesund*, <āuġin brauuᵊe> *Augenbraue*, <zibin> *sieben*	**n**	ن
/w/: <wasser> *Wasser*	**w**	و
/e/ [ɛ] (im Inlaut): <šperber> *Sperber*, <fet> *fett* **/e/ [ʔɛ-]** (im Anlaut): <anteriḫ> *Enterich*, <as> *es*, <altrny> *Eltern* **/e/ [(.)ʔɛ-]** (im Anlaut und im Silbenonset): <ᵊelbuġin> *Ellbogen*, <fir ᵊeqkiġ> *viereckig* **/e/** (im Anlaut, bes. vor *-r-* [im Verbpräfix *er-*]): <ᵊrwartin> *erwarten*, <ᵊrwaḫin> *erwachen*, <ᵊrlrnin> *erlernen*, <ᵊrmudin> *ermüden*, <ᵊrinnrny> *erinnern*, <ᵊr> *er* **/e/** (in Silbe onset): <beᵊndiġin> *beendigen*	**e** **a** **ᵊe** **ᵊ** **ᵊ**	ه ا ئه ئ ئ

[17] Vgl. im Standarddeutschen gibt es [s-] im Anlaut vor Vokalen nicht.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Deutsche Phoneme (mit ihren Allophonen)	**Trans-litera-tion**	**Arabische graphemati-sche Entspre-chung(en)**
/e/ [ə] (in unakzentuierten Silben, bes. im Auslaut): <qauen> *kauen*, <tauen> *tauen*, <sardine> *Sardine* **[ə]** (in unakzentuierten Silben): <āuġin brauuᵊe> *Augenbraue*, <maġin> *Magen*, <zibin> *sieben*, gewöhnlich in der Infinitivendung-*en* /-ən/ [-ən], [-n̩]	**e** **i**	ه ي
/ē/ (im Inlaut): <qamel> *Kameel* (mod.nhd. *Kamel*), <leber> *Leber*, <kele> *Kehle*	**e**	ه
/ǟ/ [ɛː] : <šedel> *Schädel*	**e**	ه
/i/: <diq darmy> *Dickdarm* **/i/**: (selten) <mynister> *Minister*, <auffyčir> *Offizier* **/i/**: (im Anlaut): <aist> *ist*, <airrin hauz> *Irrenhaus*	**i** **y** **ai**	ي ى اي
/ī/: <zibin> *sieben* **/ī/** (selten): <lybin> *lieben*, <wy> *wie*, <zy> *sie*	**i** **y**	ي ى
/j/ (als zweiter Teil von Diphthongen): <zain> *sein* **/j/**: <iaqke> *Jacke*, <iunġ> *jung* **/j/** (als zweiter Teil von Diphthongen [selten]): <šraybin> *schreiben*, <ārbaytin> *arbeiten*	**i** **y** **y**	ي ى ى
/u/ (im Inlaut): <mutter> *Mutter*, <hund> *Hund* **/u/** (im Anlaut): <aunzr> *unser* **/u/** (im Anlaut): <ᶜund> *und*, <ᶜunterhuze> *Unterhose*, <ziḫ ᶜuntrhaltin> *sich unterhalten*, <ᶜum> *um*, <ᶜum ārmin> *umarmen*	**u** **au** **ᶜu**	و او عو
/ū/ (im Inlaut): <hun> *Huhn* **/ū/** (im Anlaut): <ᶜur ġrus fater> *Urgroßvater*, <ᶜur ġrus mutter> *Urgroßmutter*	**u** **ᶜu**	و عو
/o/ (im Inlaut): <tuḫtr> *Tochter*, <ārtišuqke> *Artischocke* **/o/** (im Anlaut): <aunqel> *Onkel*, <auffyčir> *Offizier*	**u** **au**	و او
/ō/ (im Inlaut): <zun> *Sohn*, <huze> *Hose* **/ō/** (im Anlaut): <aur> *Ohr*, <aubr ārč> *Oberarzt*	**u** **au**	و او
/ü/: <muqke> *Mücke*	**u**	و
/ǖ/ (im Inlaut): <zus> *süß*, <tur> *Tür* **/ǖ/ [ʔʏː-]** (im Anlaut): <aubrmurġn> *übermorgen* **/ǖ/ [ʔʏː-]** (im Anlaut): <ᶜuberruq> *Überrock* **/ǖ/ [ʔʏː-]** (im Anlaut): <ᵊubunġ> *Übung*	**u** **au** **ᶜu** **ᵊu**	و او عو ئو
/ö/ (im Inlaut): <šupfer> *Schöpfer* **/ö/ [ʔœ-]** (im Anlaut): <ᵊufnin> *öffnen*	**u** **ᵊu**	و ئو
/ȫ/ (im Inlaut): <fugel> *Vögel*, <šun> *schön* **/ȫ/ [.ʔøː-]** (im Silbenonset): <aulifin ᵊul> *Olivenöl*	**u** **ᵊu**	و ئو
/a/ (im Inlaut): <wasser> *Wasser*, <qalb> *Kalb* **/a/** (im Auslaut): <sufa> *Sofa* **/a/ [(.)ʔa-]** (im Anlaut oder im Silbenonset): <āffe> *Affe*, <ārbaytin> *arbeiten*, <ālt> *alt*, <ᶜum ārmin> *umarmen*	**a** **ā**	ا آ

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Deutsche Phoneme (mit ihren Allophonen)	**Trans-litera-tion**	**Arabische graphematische Entsprechung(en)**
/ā/ (im Inlaut): <han> *Hahn*, <haze> *Hase*, <suldat> *Soldat* **/ā/ [(.)ʔaː-]** (im Anlaut oder im Silbenonset): <ābnd> *Abend*, <zun ābnd> *Sonnabend*, <lineāl> *Lineal*	**a** **ā**	ا آ
/au/ (im Inlaut/Auslaut): <pfau> *Pfau*, <qranqn hauz> *Krankenhaus* **/au/** (im Inlaut und Auslaut): <blauu> *blau*, <bauuḫ> *Bauch* **/au/ [ʔaʊ-]** (im [akzentuierten] Anlaut): <āuuster> *Auster*, <āuusgehin> *ausgehen* **/au/ [ʔaʊ-]** (im [unakzentuiertem?] Anlaut): <āuġust> *August*, but <āuge> *Auge*, <āuġin ārč> *Augenarzt*	**au** **auu** **āuu** **āu**	او اوو آوو آو
/oi/ [ʔɔɪ-] (im Anlaut): <auier> *euer*, <auirer> *eurer*, <auiḫ> *euch* **/oi/** <eu>: <tuyfel> *Teufel* **/oi/** <äu>: <bruytiġām> *Bräutigam* **/oi/** <eu>: <duič> *deutsch*, <huite> *heute* **/oi/** <nuiun ʿund nuiun čiġ> *neunundneunzig*	**aui** **uy** **uy** **ui** **uiu**	اوي وى وى وي ويو
/ai/ [ʔaɪ-] (im Anlaut): <āis> *Eis* **/ai/** (im Inlaut): <laiḫt> *leicht*, <qraide> *Kreide* **/ai/** (im Inlaut): <blay feder> *Bleifeder*, <blay štift> *Bleistift*, <šun šray bin> *schönschreiben*, <drayčen> *dreizehn* **/ai/** (in Inlaut): <drayisiġ> *dreißig* **/ai/** (im Auslaut): <ārčnay> *Arznei*, <dray> *drei*, <bay> *bei*, <zay> *sei* **/ai/** (im Auslaut, selten) <drāy> *drei*	**āi** **ai** **ay** **aii** **ay** **āy**	آي اي اى ايي اى آى

4.3 Weitere ‚Schreibregeln'

4.3.1 Geminaten

4.3.1.1 Wortinterne deutsche orthographische Geminaten werden gewöhnlich durch arabische Geminaten nachgeahmt: <auffyčir> *Offizier*, <airrin hauz> *Irrenhaus*, <ʾessn> *essen*. Das arabische Zeichen für die Gemination, *damma*, wird nicht gebraucht, nicht einmal im Namen des Propheten, <Muḥammed> *Mohammed*.

4.3.1.2 Deutsche geschriebene Geminaten im Auslaut oder der Silbencoda werden nicht als Geminaten geschrieben: <ʾelbuġin> *Ellbogen*, <mitwuḫ> *Mittwoch*, <fet> *fett*, <bt> *Bett*.

4.3.1.3 Deutsch <ck> wird im Inlaut zwischen Vokalen arab. <qk> geschrieben, wahrscheinlich in der Absicht, den deutschen Digraphen zu imitieren: <ārtišuqke> *Artischocke*, <(fr)šiluqkin> *(ver)schlucken*, <fir ʾeqkiġ> *viereckig*.

4.3.1.4 Im Morphem-Auslaut oder Wort-Auslaut wird dt. <ck> als <-q>, also phonologisch wiedergegeben: <diq darmy> *Dickdarm*. Somit folgt die Schreibung hier der Regel (4.3.1.2), dass Geminaten im Auslaut oder der Silbencoda nicht als Geminaten geschrieben werden.

4.3.1.5 Als Unterklasse der Geminaten kann das Graphem <ß> angesehen werden, das aus dem Cluster <ſʒ> hervorgegangen ist, wie man an der gotischen Form des Graphems noch klar erkennt. Während (silben-)schließendes *-ß*./# generell gemäß seiner Aussprache als <-s> erscheint (man vgl. auch wie auslautende deutsche Geminaten sonst geschrieben werden [4.3.1.2]), wird mittleres <-ß-> gewöhnlich als <-ss-> und einmal als <-s-> geschrieben:

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

<ġrus> *groß*, <hais> *heiß vs.* <flaissiġ> *fleißig*, <messiġ> *mäßig*, <waisse> *weiße vs.* <drayisiġ> *dreißig*.

4.3.2 *Cw*-Cluster

<CiwV-> wird für #/.*CwV*- geschrieben: <čiwibel> *Zwiebel*, <šiwiger mutter> *Schwiegermutter*, <šiwert fiš> *Schwertfisch*, <šiwarč> *schwarz*, <šiwarčes> *schwarzes*, <fršiwindin> *verschwinden*, <čiway> *zwei*.

Ausnahmen sind selten: <šwangr> *schwanger*, <čway> *zwei*.

4.3.3 *Cl*-Cluster

<CilV-> wird für #/.*ClV*- geschrieben: <šileḫt> *schlecht*, <(fr)šiluqkin> *(ver)schlucken*, <šilaḫt> *Schlacht*.

Ausnahmen sind selten: <šlange> *Schlange*.

4.3.4 Andere Cluster

Vereinzelt werden auch andere Cluster gemieden bzw. aufgelöst: <zamistaġ> *Samstag*.

4.3.5 Die Schreibungen von /e/

4.3.5.1 Kurzes dt. /e/ wird im Auslaut bzw. in der Silbencoda (bes. vor /r/) bzw. in der Endung / dem Suffix -*er* (gesprochen [-ɐ]) meist nicht geschrieben:

4.3.5.1.1 Im Suffix -*er*: <tuḫtr> *Tochter*, <fenstr> *Fenster*, <šifr> *Schiefer*, <messr> *Messer*, <šulr> *Schüler*, <tur hutr> *Türhüter*, <qust šulr> *Kostschüler*, <mdičinr> *Mediziner*, <dulmečr> *Dolmetscher*.

Einige Ausnahmen mit <-e-> sind: <mutter> *Mutter*, <šiuiger mutter> *Schwiegermutter*, <lerer> *Lehrer*, <šmettrliṅġ> *Schmetterling*, <ᶜuntrriḫt> *Unterricht*.

4.3.5.1.2 In der Endung -*er*: <aimmr> *immer*, <waitr> *weiter*, <mainr> *meiner*, <dainr> *deiner*, <zainr> *seiner*, <aunzr> *unser*.

Einige Ausnahmen mit <-e-> sind: <āber> *aber*, <aunzer> *unser*, <auier> *euer*, <auirer> *eurer*, <airer> *ihrer*.

Auffällig sind Dubletten wie <aunzer> *unser* vs. <aunzr> *unser*.

4.3.5.2 Aber im Auslaut in Einsilblern, wo -*er* als [-eɐ] ausgesprochen wird, wird es manchmal geschrieben:

Artikel Nom.Mask.Sg. <dr> *der* (aber manchmal auch <der>), <wer> *wer*, <hr> *Herr*.

Eine Ausnahme ist <zr> *sehr* [zeːɐ], wo der Langvokal /ē/ nicht geschrieben wird.

4.3.5.3 Im Suffix -*el*: <čirkl> *Zirkel*, <štat firtl> *Stadtviertel*, <mittl> *Mittel*.

4.3.5.4 In der Endung -*en* (Infinitive, Substantive, Adjektive im Sg. oder Pl.):

<lrnn> *lernen*, <frkaufn> *verkaufen*, <qaufn> *kaufen*, <bittn> *bitten*, <belebn> *beleben*, <quḫn> *kochen*, <ārbaitn> *arbeiten* – <ġartn> *Garten*, <aufn> *Ofen*, <brun nn> *Brunnen* – <weġn> *wegen*, <murġn> *morgen* – <ġutn> *guten*.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Bei <widerzehn> *Wiedersehen* ist entweder das in der Bühnenaussprache vorhandene /ə/ in der Endung nicht geschrieben oder es wird die in gesprochener Rede übliche Aussprache [ˈviːdɐzeːn] wiedergegeben.

Am Ende des ersten Glieds eines Kompositums: <qranqn hauz> *Krankenhaus*, <qnabn šule> *Knabenschule*.

Ausnahmsweise werden solche Wörter mit <-e-> geschrieben: <qauen> *kauen*, <zitsen> *sitzen*.

Ebenfalls selten ist die Schreibung dieses Morphems mit <-in>: <airrin hauz> *Irrenhaus*).

4.3.5.5 Die Infinitivendung *-eln* wird in etwa gleichmäßig mit und ohne <-e-> geschrieben: <zammlny> *sammeln* vs. <šuttelny> *schütteln*.

4.3.5.6 Gen.-Sg.-Endung *-es*: <ġutas> *Gottes*.

4.3.5.7 In den Präfixen *er-*, *ver-*: <frkaufn> *verkaufen*, <frberġin> *verbergen*, <ᵊrlrnin> *erlernen*, <ᵊrinnrny> *erinnern*.

4.3.5.8 Im Inlaut vor /r/: <lrnn> *lernen*, <nrrin> *Närrin*, <čan šmrč> *Zahnschmerz*, <quḫ hrd> *Kochherd*, <krbel> *Kerbel*, <ziḫ ᶜuntrhaltin> *sich unterhalten*, <pfrd> *Pferd*, <pfrde ārč> *Pferdearzt*, <nrf> *Nerv*, <črtliḫ> *zärtlich*, <geštrn> *gestern*, <furgeštrny> *vorgestern*.

4.3.5.9 In anderen Positionen: <wste> *Weste*, <lqtiun> *Lektion*, <mdičin šule> *Medizinschule*, <aurgehnge> *Ohrgehänge*, <bt> *Bett*, <tller> *Teller*, <hft> *Heft*, <geburst hlfr> [sic] *Geburtshelfer*, <raiznde> *Reisende*, <zun ābnd> *Sonnabend*, <ābnd> *Abend*.

4.3.5.10 Ausnahmsweise werden auch /ē/, /ā̄/ vor /r/ nicht geschrieben: <militr šule> *Militärschule*, <mdḫn šule> *Mädchenschule*, <mr> *mehr*, <zr> *sehr*, <wuhr> *woher*.

4.3.5.11 Ausnahmsweise werden /ē/, /ā̄/ vor /n/ nicht geschrieben: <ātn> *Athen*.

4.3.5.12 Ausnahmsweise wird /i/ nicht geschrieben: <Ḫrstus> *Christus*.

4.3.6 Weitere allgemeine Regeln

4.3.6.1 Morpheme werden nicht notwendigerweise in Simplizia und Komposita oder Ableitungen in identischer Schreibweise geboten: <ārtiġ> *artig* vs. <ᶜunartiġ> *unartig*; <āuge> *Auge* vs. <āuġin wimper> *Augenwimper*, <āuġin lid> *Augenlid*; Pers.Pron. <ᶜunz> *uns* vs. Poss.Pron. <aunzr> *unser*.

4.3.6.2 Generell gibt es eine Tendenz, die deutsche Orthographie nachzuahmen: <qatče> *Katze*.

4.3.6.3 Die Transkription der deutschen Wörter scheint nach der im Text vorliegenden Form erfolgt zu sein. Wie man sich das konkret vorzustellen hat, ist unklar: Möglicherweise lagen die Schreibfehler in den deutschen Wörtern bereits im Manuskript bzw. in einer nur die deutschen Wörter in Originalschrift enthaltenden gedruckten Version vor. Eine Kontrolle anhand eines Wörterbuchs des Deutschen scheint jedenfalls unmittelbar vor der Drucklegung nicht mehr erfolgt zu sein. Folglich wurden entweder falsch geschriebene oder bereits zu jener Zeit archaische Wortformen transkribiert, was für den türkischen Nutzer eine Aussprache der Wörter nahelegte, die teils wenig mit dem tatsächlichen Wort zu tun hatte:

S. 14: <čurtiġ> *zortig* statt *zornig*. Hier werden im Arabischen <n> und <t> im Wortinnern nur durch die Zahl der Punkte über dem Grundgraphem differenziert (ein Punkt bei <n>, zwei Punkte bei <t>);

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

S. 19: <munče> *Münze* statt *Minze*;

S. 26: <kuhe> *Kühe* statt *Küche*;

S. 33: <bešmussin> *beschmüßen* statt *beschmutzen*;

S. 35: <blissin> *blissen* statt *blitzen*;

S. 36: <ᶜum dreḫin> *um drechen* statt *umdrehen*;

S. 39: <špečil ārč> *Speziel arzt* statt *Spezialarzt*;

S. 40: <geburst hlfr> *Gebürst helfer* statt *Geburtshelfer*;

S. 40: <hehamme> *Hehamme* statt *Hebamme*.

Zweimal wird ein deutscher Druckfehler nicht in der Transkription wiedergegeben:

S. 41: <du ġrusser ġut> *Du grotzer* (für *großer*) *Gott*;

S. 39: <wrkupft da> *Werkopt da* statt *wer klopft da*? (obwohl das <f> in der deutschen Form fehlt, ist es in der Transkription vorhanden, doch fehlt beide Male das <-l-> in *klopft*).

Entweder eine falsche Transkription auf Grundlage eines Druckfehlers oder aber eine archaische Wortform (so schrieb etwa Goethe in seinem Tagebuch der italienischen Reise 1787/88 mehrfach von *Pfirschen*) liegt vor in:

S. 18: <pfirše> *Pfirsche* statt *Pfirsiche*.

4.3.6.4 Schreibungen, die unmittelbar der deutschen Schreibung folgen:

S. 40: <passagir> *Passagier* [pasaˈʒiːɐ].

4.3.6.5 Auslautendes /-rn/, /-rm/ wird gewöhnlich <-rny>, <-rmy> geschrieben:

bei Verben: <ᵊrinnrny> *erinnern*, <hagelny> *hageln*, <dunnrny> *donnern* <zammlny> *sammeln*, <šuttelny> *schütteln*;

bei Adverbien: <furgeštrny> *vorgestern* vs. <geštrn> *gestern*;

bei Substantiven und Adjektiven: <harny> *Harn*, <gehirny> *Gehirn*, <mušelny> *Muscheln*, <ārmy> *arm*, *Arm*, <warmy> *warm*, <diq darmy> *Dickdarm*, <zaidinwurmy> *Seidenwurm*.

Eine Ausnahme oder ein Druckfehler leigt vor bei <yuny> *Juni*.

4.3.7 Homographen

Es gibt ein Homographenpaar (das aber keine homophonen Wörter wiedergibt):

<ġut> *gut* [guːtʰ] vs. <ġut> *Gott* [gɔtʰ].

4.3.8 Wörter, die Schreibvarianten zeigen

Folgende Wörter legen nahe anzunehmen, dass Scherefeddin Pascha sein Werk nicht mit einem durchdachten Plan unternommen hat.

4.3.8.1 Anlautvariation: <ᵊr> : <ar> *er*, <ᵊs> : <as> *es*, <čiway> : <čway> *zwei*.

4.3.8.2 Inlautvariation: <dain> : <dāin> *dein*, <nuyun> : <nuiun> *neun*.

4.3.8.3 Auslautvariation: <dr> : <der> *der*, <aunzr> : <aunzer> *unser*, <drāy> : <dray> *drei*, <ainen> : <ainn> *ihnen*, <ġrusher> : <hr> *-herr*, *Herr*.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

5. Zusammenfassung

Die Analyse der Schreibgewohnheiten hat gezeigt, dass Scherefeddin Pascha es nicht geschafft hat, ein eindeutiges System der Transkription für das Deutsche zu entwickeln. Weder bei den Vokalen noch bei den Konsonanten hat er für mehr als einzelne Laute eine unzweideutige Phonem-Graphem-Entsprechung gewählt. Er nutzte weder die grundsätzlich bestehenden Möglichkeiten der arabischen Schrift und ihrer Diakritika noch die orthographischen Traditionen der Verschriftung des Osmanisch-Türkischen aus, um die Lautgestalt der deutschen Wörter adäquat und eindeutig wiederzugeben. Er konnte auf kein bestehendes Transkriptionssystem zurückgreifen (alle älteren Transkriptionsversuche blieben ebenfalls Einzelerscheinungen), konnte aber auch selbst keines mehr etablieren – mit der Einführung der Lateinschrift 1928 entfiel eine solche Notwendigkeit ohnehin.

Literatur

Arslan, M. (2011): Menâzımü'l-Cevâhir: Arapça-Farsça-Türkçe Manzum Sözlük. Ankara.

Arslan, A. (2016): Tuhfe-i Hüsâmî: İnceleme, Çeviri Yazılı Metin, Dizin. M.A. Thesis, Eskişehir Osmangazi University, Institute of Social Sciences.

Bairašauskaitė, T. (2021): Lietuvos totorių istorija. Vilnius.

Bichlmeier, H. (2020): Writing German with Arabic script in ca. 1590 and in 1916: a comparison. In: Linguistica Lettica 27, 2019 [2020], S. 143–160.

Bichlmeier, H./Doğan Averbek, G. (2021): Chapter two: linguistic features und literary value of *Almanca Tuhfe / Deutsches Geschenk* (1916). In: Van de Velde, H./Dolezal, F. (Hg.): Broadening perspectives in the history of dictionaries und word studies. Cambridge, S. 31–57.

Bichlmeier, H./Ivušić, B. (2013): Zur dialektologischen Einordnung der deutschen Texte einer osmanischen Sammelhundschrift vom Ende des 16. Jahrhunderts. In: Harnisch, R./Graßl, S./Spannbauer-Pollmann, R. (Hg.): Strömungen in der Entwicklung der Dialekte und ihrer Erforschung. Beiträge zur 11. Bayerisch-Österreichischen Dialektologentagung in Passau, September 2010. Regensburg, S. 365–384.

Blau, O. (1868): Bosnisch-türkische Sprachdenkmäler. Gesammelt, gesichtet und herausgegeben von O. Blau. Leipzig.

Dankoff, R./Kut, A. T./Weitenberg, J. S. (1996): The versified Armenian-Turkish glossary by Kalayi ca. 1800. Cleveland.

Doğan Averbek, G. (2018a): Dillerinden Biri Türkçe Olan Manzum Sözlükler Üzerine Yapılan Çalışmalar Bibliyografyası. In: Divan Edebiyatı Araştırmaları Dergisi 21, S. 85–114.

Doğan Averbek, G. (2018b): Türkçe-Arnavutça Manzum Sözlük Dürre-i Manzûme'nin Bilinmeyen İki Nüshası. In: FSM İlmî Araştırmalar İnsan ve Toplum Bilimleri Dergisi 12, S. 223–242.

Doğan Averbek, G./Bichlmeier, H. (2020): Almanca Tuhfe – Deutsches Geschenk (1916). The only versified Turkish-German dictionary. With an introduction on versified dictionaries as coursebooks. Berlin u. a.

Doğan Averbek, G./Bichlmeier, H. (2021): Osmanlı eğitim sisteminde ders kitabı olarak manzum sözlükler (tuhfeler) / Versified dictionaries (Tuhfe Genre) as coursebooks in the Ottoman education system. In: Divan Edebiyatı Araştırmaları Dergisi / The Journal of Ottoman Literature Studies 26, S. 61–82.

Erik, B. A. (1982): Tuhfe-i Nuriyye ve Zeyl-i Tuhfe-i Nuriyye. Graduation Thesis, Hacettepe University.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Fedayî, Mehmed Dede (2019): Tuhfe-i Fedayî. Editor G. Doğan Averbek. Istanbul.

Gragger, R. (1927a): Der magyarische Text von Murād's ‚Glaubenshymnus' mit deutscher Übersetzung. In: Babinger, F. et al. (Hg.): Literaturdenkmäler aus Ungarns Türkenzeit. Nach Handschriften in Oxford und Wien. Berlin/Leipzig, S. 55–69.

Gragger, R. (1927b): Türkisch-ungarische Kulturbeziehungen. In: Babinger, F. et al. (Hg.): Literaturdenkmäler aus Ungarns Türkenzeit. Nach Handschriften in Oxford und Wien. Berlin/Leipzig, S. 1–32.

Grüßhaber, G. (2018): The German spirit in the Ottoman army 1908–1938. A history of military knowledge transfer. Berlin.

Haber. 19 December 1926. İrtihal news.

Hadžijahić, Muhamed (1941): Aljamiado Literatura (kod Hrvata). In: Ujević, M. (Hg.): Hrvatska Enciklopedija. Encyclopaedia Croatica. Bd. I: *A-Automobil.* Zagreb, S. 300–301.

Hadžijahić, Muhamed (1955): Književnost na arabici (aljamiado). In: Babić, A./Krleža, M. et al. Enciklopedija Jugoslavije. Bd. I: *A-Bosk.* Zagreb, S. 144–145.

Ivušić, B. (2012): Die südslavischen Aljamiado-Texte der Wiener Sammelhundschrift Flügel 2006. In: Welt der Slaven 57/2, S. 380–398.

Ivušić, B. (2013): Sieben auf einen Streich. Manuskript des Monats 04/2013. https://www.manuscript-cultures.uni-hamburg.de/mom/2013_04_mom.html (20.06.2021).

Kaya, H./Ayçiçeği, B. (2019): Müsellesnâme: Osman Şâkir'in Manzum Sözlüğü. Istanbul.

Kılıç, A. (2007): Türkçe-Arapça Manzum Sözlüklerden Sübha-i Sıbyân, 2: Metin. In: Turkish Studies 2 (1), S. 29–71.

Kontzi, R. (1980): Aljamiado-Literatur. In: Lexikon des Mittelalters I: Aachen bis Bettelsordenskirchen. Stuttgart, S. 415.

Miškienienė, Galina et al. (Hg.) (2009): Ivano Luckevičiaus Kitabas. Lietuvos totorių kultūros paminklas. Vilnius.

Mittwoch, E./Mordtmann, J. H. (1927): Die Wiener Sammelhundschrift. In: Babinger, F. et al. (Hg.): Literaturdenkmäler aus Ungarns Türkenzeit. Nach Handschriften in Oxford und Wien. Berlin/Leipzig, S. 70–87.

Mittwoch, E. (1927): Die deutschen, magyarischen, kroatischen und lateinischen Texte der Wiener Sammelhundschrift. In: Babinger, F. et al. (Hg.): Literaturdenkmäler aus Ungarns Türkenzeit. Nach Handschriften in Oxford und Wien. Berlin/Leipzig, S. 88–130.

Muhtar, C. (1993): İki Kur'an Sözlüğü: Luğat-ı Ferişteoğlu ve Luğat-ı Kânûn-i İlâhî. Istanbul.

Neweklowsky, G. (1996): Die bosnisch-herzegowinischen Muslime – Geschichte, Bräuche, Alltagskultur. Klagenfurt/Salzburg.

Römer, C. (2014): Cultural assimilation of a 16th-century new muslim. The Mecmua ÖNB A. F. 437. In: Şahin, İ./Egawa, H./Erdoğan Özünlü, E./Öğün, T. (Hg.): CIÉPO 19. Osmanlı Öncesi ve Dönemi Tarihi Araştırmaları. 2 Bde. Istanbul, S. 607–619.

Rossi, E. (1946): Notizia su un manoscritto del canzoniere di Neẓīm (secolo xvii-xviii) in caratteri arabi e lingua albanese. In: Rivista degli studi orientali 21/2–4 (November), S. 219–246.

Scherefeddin P. (1916): Almanca tuḥfe [Ālmānje tḥfe]. Deutsches Geschenk. Constantinopel (= Istanbul 1332 AH).

Siebs, Th. (1905): Deutsche Bühnenaussprache. 3. Auflage. Berlin/Köln/Leipzig.

Siebs, Th. (1915): Deutsche Bühnenaussprache. 11. Auflage. Bonn.

Staatsarchiv der Republik Türkei, 253–257, Y.PRK.ASK.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Turan, M. (2012): Hasan Rızâyî ve Kân-ı Maʿânî İsimli Manzum Sözlüğü. In: Turkish Studies 7 (4), S. 2939–2992.

Uskufi, M. H. (2001): Makbūl-i ʾArif: Potur Šahidiya. Tuzla.

Vehbî, S. (2012): Tuhfe-i Vehbî: Metin, Dizin, Tıpkıbasım. Hrsg. v. A Yenikale. Kahramanmaraş.

Verburg, A. C. (1997): The Tuhfe-i Şāhidī: a sixteenth century Persian-Ottoman dictionary in rhyme. In: Archivum Ottomanicum XV, S. 5–87.

Yıldırım, N. (2006): Sağlık Kuruluşları. In: Surların Öte Yanı: Zeytinburnu. 3. Auflage. Istanbul, S. 296–355.

Yurtseven, N. (2003): Türk Edebiyatında Arapça-Türkçe Manzum Lügatlar ve Sünbülzade Vehbi'nin Nuhbe'si. M.A. Thesis, Ankara University, Institute of Social Sciences.

Yûsuf Hâlis (2006): Miftâh-ı Lisân: Manzum Fransızca-Türkçe Sözlük. Hrsg. v. M. Kırbıyık. Istanbul.

Kontaktinformationen

Harald Bichlmeier
Sächsische Akademie der Wissenschaften zu Leipzig,
Arbeitsstelle Jena: Etymologisches Wörterbuch des Althochdeutschen
harald.bichlmeier@uni-jena.de

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

María Pozzi

DESIGN OF A DICTIONARY TO HELP SCHOOL CHILDREN TO UNDERSTAND BASIC MATHEMATICAL CONCEPTS

Abstract This paper presents the decisions behind the design of a maths dictionary for primary school children. We are aware that there has been a considerable problem regarding Mexican children's performance in maths dragging on for a long time, and far from getting better, it is getting worse. One of the probable causes seems to be the lack of coordination between maths textbooks and teaching methods. Most maths textbooks used in primary schools include lots of activities and problem-solving techniques, but hardly any conceptual information in the form of definitions or explanations. Consequently, many children learn to do things, but have difficulty understanding mathematical concepts and applying them in different contexts. To help solve this problem, at least partially, the project of the dictionary was launched aiming at helping children to grasp and understand maths concepts learned during those first six years of their formal education. The dictionary is a corpus-based terminographical product whose macrostructure, microstructure, typography, and additional information were specifically designed to help children understand mathematical concepts.

Keywords children's specialised lexicography; corpus-based terminography; mathematical terms; children's vocabulary; conceptualisation

1. Introduction

Form an early age, children start to acquire mathematical and geometrical concepts, such as those of basic shapes, quantities, distance, and to make comparisons between objects[1] within their everyday life activities. They will refine these concepts when they begin their primary school education which, in turn, will prepare them to learn more abstract concepts in a formal way. For Cabré (1993, pp. 222 f.) the main differences between terms and general language words (GLW) are: a) terms are explicitly learned while GLW are spontaneously learned in everyday life experiences; b) terms have a referential function, GLW mainly have a conative, an emotive or a phatic function; c) terms are used in one or more specific domains of knowledge, GLW are used in everyday life situations; d) terms are used in formal communicative situations, GLW are used in less formal communicative situations; e) terms are used in a professional or scientific discourse, GLW are used in general discourse; and f) terms are used by specialists, GLW are used by every speaker of a language.

Pozzi (2016, pp. 112 f.) expresses some concern in relation to the way in which linguistic units designating basic mathematical concepts are considered in practice: Are these terms or GLW? The answer depends on who is replying to the question. For an educated adult or a specialist these linguistic units probably are general language words, but for children who must learn them, teachers who must teach them, textbook authors, education policy makers and others working in related fields, these are terms, since they satisfy all but one (children are no specialists) of Cabré's criteria for being terms. Because of the lack of consensus on their nature or perhaps because these units share most of their characteristics with both

[1] Such as "*a* is bigger than *b*", "*a* is smaller than *b*", "*a* is equal to *b*", "*a* and *b* are different".

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

GLW and terms, the maths vocabulary learned by children in primary school has not received a lot of attention from terminologists and linguists alike. For the purpose of this project, linguistic units representing mathematical concepts at the lowest level of specialisation are assumed to be terms.

But one thing is for certain: from that age on, children will eagerly learn and like maths depending on how well they understand and make their own these first concepts taught at school. Thus, the great importance of acquiring both concepts and terms right from the time of their first encounter.

2. Background

After several years of planning, discussing, and arguing about pros and cons, some sixty odd years ago, Barriga (2022) tells us, Mexico made a major political move to provide "a public, universal and free education to open the paths to the liberating hope that knowledge can offer". In practical terms, this meant a standardised national curriculum, and a set of free-for-all textbooks (LTG)[2] for each child attending primary school whether urban or rural, public or private. This, in principle, was an extraordinary step to guarantee the same education for every child. However, from the very first edition, the LTG have had a high degree of controversy in almost every aspect, including the quality of their content, the constant changes, and other factors that have affected children's learning. For teachers, in turn, it has not been easy either, since, for every new edition they have had to face the confusion of constant and abrupt changes in approach and terminology, and many teachers have not been able to implement those changes in a satisfactory way.

The last few editions of maths textbooks have promoted the constructivist approach in which, following Piaget's (1967) ideas, children are expected to discover principles and construct their own understanding of concepts based on observation and on "doing things". Vygotsky's (1978) idea of "knowledge is constructed in a social context before it is acquired" has also been implemented. According to Pozzi (2016, p. 113), this change of approach has had a considerable effect on the teacher-children relationship in the classroom and on the content and presentation of textbooks. Teachers are expected to provide the appropriate guidance and environment to students so that they can make their own observations and produce their own mental constructs.

This new approach, although welcome in principle, has not proved effective for several reasons, among which, in my opinion, the following three stand out:

- Teachers were not offered compulsory training on the constructivist approach to teaching and learning; some of them are still oblivious to these changes.
- Although textbooks follow the constructivist approach, most schools do not conduct teacher assessment to ensure they provide the necessary guidance to students to help them conceptualise and construct their own knowledge; consequently, there is an obvious incompatibility between teaching methods and textbooks.

[2] From its initials in Spanish, *libros de texto gratuito* (LTG).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

– Maths textbooks contain activities, exercises, illustrations, problems to be solved, etc., but hardly any explicit conceptual information in the form of definitions or explanations.

As a result, if children fail to understand what they are doing, they cannot conceptualise well and in turn, the following concepts they are supposed to learn, based on the ones they could not understand, will not be understood either. This process can go on and on … In this context, it is not surprising that many children do not like maths and cannot progress as well as they should.

Perhaps this lack of conceptual understanding could be one of the key factors for the disappointing result in the Programme for International Student Assessment (PISA) conducted by the Organisation for Economic Co-operation and Development (OECD) that measures the level of achievement of children aged 6 to15 years, in mathematics, science and reading skills. According to the latest PISA assessment (OECD, 2019): "In Mexico […] the mean score in PISA performance in science and mathematics is well below average". It came out in 64th place out of a total of 79 countries in maths, with a score of 409 points, 182 less than the highest of 591 obtained by China, and only 82 points ahead of the lowest of 325 obtained by the Dominican Republic. In addition, if we consider that, according to the International Monetary Fund (2021), Mexico is the world's 15th largest economy, these results are alarming from a social, educational, and even political and economic perspectives.

With this background in mind, I believe we can help children to improve their skills, to conceptualise and to acquire those very first mathematical concepts – and terms – that are the basis on which they will construct their own mathematical knowledge in the years to come, by strengthening their concept understanding once they have made initial observations and have had hands-on practice on a particular concept or set of related concepts. For this purpose, we decided to prepare a corpus-based specialised lexicographical product that is intended to complement – not substitute – the information provided in textbooks.

3. Design of the dictionary

3.1 Preliminary decisions

As in any lexicography/terminography project, the product's overall design and the methodology to be followed were defined based on the answers to the following three basic questions:

a) *Why is it needed?* To help primary school children to improve their understanding of basic mathematical concepts.

b) *Who will benefit from it?* Primary school children, teachers, parents, and anyone whose understanding of basic maths needs to be enhanced.

c) *What information will it provide and how this information will be presented?* By publishing a corpus-based *dictionary*[3] providing conceptual, linguistic, and practical information in a clear, interesting, and in an appropriate language for children.

[3] Strictly speaking, our product is not a proper dictionary, nor is it an encyclopaedia, but it is somewhere in between. We decided to call it *dictionary* because it is the simplest term.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

The *why?* and *for whom?* need no further explanation. It is the answer to the third question that accounts for the main decisions taken to implement the dictionary the way we did. And so, I now provide a succinct justification for these decisions.

Firstly, we had to admit that, although we had extensive experience in standard terminography work, we had none in children's specialised lexicography. Therefore, we had to study as many children's specialised dictionaries as we could get hold of, in Spanish, English, and French.[4] We immediately realised that writing for children is different from writing for teenagers and adults. So, we decided to conduct some user studies involving children as the dictionary's main target users.

Secondly, we designed a series of three user studies to determine: a) the degree of understanding of maths concepts, b) whether the children could illustrate the dictionary, and c) what was the most efficient way to write and present the definitions.

a) The study to determine the degree of understanding of maths concepts was performed in an informal way so that the children would feel at ease. All children,[5] from 1st to 6th grade, participated. We had a list of concepts they should have learned that year and asked who could explain one or more. Most concepts were explained in a satisfactory manner; however, it was always the same top of the class students who answered, those who are good at maths and enjoy the subject. To assess the overall situation, we went further ahead and asked each child if he or she could explain a given concept to the rest of the class. Here, the results varied a great deal. In average, two out of five admitted that they had difficulties understanding that concept and therefore, could not explain it; about 2 in 5 said they thought they understood the concept but were unable to explain it; only one in five understood the concept and could explain it.

b) In line with the current trend of presenting children's dictionaries illustrated by children (Estopà, 2021), we asked the three schools involved in the dictionary project whether the children could provide the illustrations. It was agreed that all 4th to 6th graders would participate, and it would be done during school hours. Children were randomly assigned three concepts they were already familiar with, together with their definitions and a simple picture or drawing, just to give them an idea of what was expected of them to produce, though they could illustrate the concept in whichever way they wanted. Unfortunately, these illustrations could not be used because they did not have the required quality. So, as a last resort, we decided to have them made by a professional illustrator.

c) It was evident from the beginning that, for the dictionary to succeed in its main objective, the most important data category we had to get right was the definition. To achieve this, once again, we designed a user study involving children of 4th and 6th grades. After analysing several models to define terms found in children's dictionaries, we selected the three most suitable and drafted a definition for twenty concepts for each school grade (4th and 6th) in each of the following models:

[4] Children's dictionaries we took some ideas from are listed in the References section.

[5] The user studies were carried out in three primary schools between 2018 and 2019: *Colegio Madrid*, a co-educational private school, each grade has four groups with an average of 30 pupils per group; *Colegio Oxford para niñas*, an all-girls private school, each grade has three groups with an average of 15 pupils per group; and *Colegio Oxford para niños*, an all-boys private school, each grade has three groups with an average of 15 pupils per group.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

– traditional intensional definition (superordinate concept followed by delimiting characteristics) written in words they were familiar with:

> parallel lines
>
> Lines that are the same distance apart no matter how long they are, and they never cross each other.

– definition introduced in context, in the style of the Collins Cobuild Dictionary (Sinclair 1995):

> parallel lines
>
> If two lines are **parallel**, they are the same distance apart from each other all the way along their length.

– List of essential characteristics, presented as a bulleted list, each one written as a short sentence:

> parallel lines
>
> - lines that stay the same distance apart along their whole length
> - they never cross
> - they can be curved or straight
> - they do not need to be the same length as each other

Although the content of each definition compared with the other two is very similar, the form in which each one is presented makes it easier or more difficult to grasp. To determine which model was preferred by the children, we gave each child a set of five concepts, selected at random from the twenty that had previously been defined in the three models. They had to select the order in which they found it easier to understand the definition and explain why. For both age groups the result was clear: they felt more comfortable and understood the definition better when it was presented to them as a bulleted list of characteristics written in short phrases closely followed by those presented in context. We used one or the other, depending on the nature of the concept being defined. Very few children selected the traditional intensional model for definitions in first place; therefore, we did not use it.

Once these three user studies were completed, we were able to decide the dictionary structure (macro- and microstructure) as well as its typographical design.

3.2 Macrostructure

3.2.1 The dictionary as a corpus-based project

It was clear from the beginning that the macrostructure of the dictionary should include every term representing a mathematical concept that children learn during their primary school years. To make sure all terms were identified, the dictionary had to be a corpus-oriented project.

The first step consisted in setting up an appropriate corpus. For that purpose, we had already built a larger multi-purpose annotated corpus of about 6 million tokens, the *Corpus of Basic Scientific texts in Mexican Spanish* (COCIEM), containing complete maths and science textbooks that fully cover the school curricula for students aged 6 to 18 years. For the primary maths dictionary, we extracted the subset of COCIEM containing all official maths and exercise textbooks, freely provided by the Ministry of Education for every child attending

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

any public or private primary school. These were complemented with nine additional maths textbooks corresponding to the most widely available books used in private primary schools. This ***Co**rpus of **P**rimary **S**chool **Math**s Texts* (COPSMAT) was the starting point to obtain the base list of terms for the maths dictionary's macrostructure (Pozzi 2016). It has 375,142 tokens and 8,533 types. Since the maths corpus is a subset of COCIEM, the lexical and terminological markup was already in place. Most of the terms had already been identified and tagged as a result of the term extraction and validation processes applied to COCIEM. These processes are fully described in Cabrera-Diego et al. (2011) and Vivaldi et al. (2012).

3.2.2 Entry terms in the macrostructure

The final list of terms in the macrostructure of the dictionary was compiled according to the following criteria:

a) Only math terms were to be included, which was evident; however, to avoid the accidental inclusion of terms belonging to another discipline or science, it had to be explicitly stated.

b) Terms included should be statistically significant, i.e., occurred in COPSMAT with a frequency ≥ 3, and occurred in two or more textbooks. The decision to eliminate terms with frequency 1 or 2 was made together with the teachers who participated in the user studies. Their main point was that children learn by repetition, by going back to a particular topic as many times as necessary, and they also need to be able to relate a concept with similar ones, which usually cannot be done it the concept occurs once or twice only. We also eliminated terms occurring in just one textbook regardless of its frequency because it may be a term used by an individual author, but its use may not be extended.

c) Very basic maths terms considered general language words that were not automatically extracted (e.g., line, distance, time, area, plus, unit) were included manually. This was important because specialised dictionaries tend to leave out very general terms which are usually defined in general language dictionaries. In this case, it was necessary to include them because some of them represent the most basic concepts on which children will base their understanding of more complex ones in the future.

d) Incomplete term families, for example, the names of polygons (pentagon, hexagon, octagon, decagon) were completed (heptagon, nonagon). In cases such as this, we decided to include the complete family because it is likely that any or all these terms should come up in a lesson as part of an exercise or example. In addition, from the terminographical point of view, the dictionary is expected to include complete term families in accordance with good practice measures.

The final number of terms entries that constitute the macrostructure of the dictionary is 842 including synonyms, abbreviations, signs, and symbols. The entries are presented in alphabetical order.

3.2.3 Annexes containing useful and grouped data

After careful consideration, we decided to provide at the end of the dictionary an additional section containing useful information, in the form of tables, grouped data, conversion tables, formulae, 2D shapes, 3D objects, etc. The following list shows the title of each piece of information included:

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

- 2-D curve shapes
- 3-D objects: prisms, pyramids, cylinder, cone
- Abbreviations
- Addition tables
- Angles
- Celsius to Fahrenheit conversion formulae
- Days of the week
- Decimal system
- Divisibility rules
- Equivalent fraction / decimal / percentage table
- Equivalent fractions
- Formulae to calculate the area of 2D shapes
- Formulae to calculate the area of 3D objects
- Formulae to calculate the perimeter of 2D shapes
- Mathematical signs and symbols
- Metric, imperial, and American units
- Months of the year
- Multiplication tables
- Number line
- Prime numbers to 200
- Principal monetary units
- Quadrilaterals
- Regular polygons
- Roman numerals
- Square and cube roots whose result is an integer number 1 to 10
- Squared and cubed numbers to 10
- Table showing volume – capacity – mass relationship
- The circle
- Transformations: rotation, translation, reflection
- Triangles
- Units of time

3.3 Microstructure

Doubtless, the microstructure is the most important part of dictionary design, as it establishes the relationship between the author and the end user. It will succeed in its primary

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

objective if it satisfies the user's needs. Before starting the practical terminographical work, a decision had to be made on the theoretical framework we would adhere to and consequently, the methodology that would be followed to ensure the success of the dictionary.

3.3.1 Theoretical framework

Children have their first encounter with terminology when they start primary school. They should learn their first mathematical concepts together with their corresponding terms in such a way that they really grasp the concept and learn to use the term for life. If the concept is well understood, it will not be confused or forgotten, and whenever they learn a new concept based on the one they have already acquired, they will easily understand it.

Since the main motivation for the preparation of the dictionary was the specific needs of children to help them understand maths concepts, our theoretical starting point was the principle of appropriateness proposed by Cabré in her Communicative Theory of Terminology (1999, p. 137), by which she states that

> [...] each specific job adopts a strategy based on its subject matter, objectives, context, elements involved and available resources. The methodology, therefore, [...] adapts itself to the circumstances without contravening the principles; the methodological appropriateness is paramount.

Although conceived as a concept-oriented product, the dictionary also provides entries for synonyms to make it possible for the children to consult the dictionary by whichever term they need to look for. For the online edition this will be transparent to the user because all synonyms, variants, and abbreviations will automatically redirect the search to the main entry term record.

In this context, the microstructure of the dictionary was set to include the following broad data categories: a) term-related data, b) concept-related data, and c) practical information data.

3.3.2 Term-related data

Term-related data associated with the linguistic information are important because they provide the primary access to the dictionary through the entry term, which can be a single or a multiword term, in the form it usually occurs in running text. These include entry term, synonyms, variants, abbreviations, signs, and symbols. The part of speech (POS) information corresponding to each term and its synonyms, variants and abbreviations is also included.[6]

When an entry term has synonyms or abbreviations, each one will have its own separate entry in the dictionary, but it will refer the user to the main entry.[7] Spelling variants are treated as synonyms.

3.3.3 Concept-related data

As the main objective of the dictionary is to help children to understand mathematical concepts, we included more concept-related data than it is traditionally provided in children's

[6] In line with this type of dictionaries, the POS of term entries is limited to nouns, adjectives and verbs.

[7] This is true for the print edition, although it will not be necessary for the online version; however, it will still refer the user to the main entry term representing the concept in question.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

dictionaries. In addition to children-oriented definitions (see section 3.1), we included the following data categories: a) maths subdomain to which the concept belongs, b) how to obtain a quantity, c) illustration, d) cross references.

a) *Subdomain* refers to the maths branch in which the concept is used: numbers and number systems; arithmetic; algebra; geometry; probability, statistics, and graphs; general vocabulary.

b) *Quantity obtention* refers to those concepts representing a quantity that can be obtained by applying a formula, as in the case of area or volume, or by means of an algorithm like solving a simple equation, adding fractions, or finding the highest common factor of a set of integers.

c) *Illustration* which shows an instance of the abstract concept and, where applicable, another instance showing one or more real life objects where the concept is applied. Sometimes, it also includes accurate diagrams and graphics that help to visualise the concept.

d) C*ross references* indicate terms included in the dictionary that are closely related to the entry term that can help children to establish explicit relations between the concepts they represent.

3.3.4 Practical and interesting data

This set of data was added to help children go from the concept to its application and vice versa. In some cases, they know how to apply a formula to solve a problem but do not understand the underlying concept. In others, they think they understand the concept but cannot apply that knowledge to solve a related problem.

These data include examples of a) how the concept is applied, b) additional information (encyclopaedic or interesting data), c) example, and d) problem solved explaining step by step how it is done.

a) *Application* refers to how the concept can be applied in real-life situations and provides an example. This is an important data category as it allows children to establish the relation between the abstract concept and its use in real life or from an object to its abstraction in a mathematical concept. For example, it goes from the concept of 'sphere' to the object beach ball and vice versa, from real life objects to the abstraction and concept formation, and mathematical description, as in the case of a snail shell to 'spirals' and 'curve description'.

b) *Additional information* refers to any information that might be interesting or useful to the children, such as some kind of encyclopaedic information, associated historical events, different uses and applications, related curious data, etc.

c) *Example* shows the use of the term in context. The term is highlighted in a different colour.

d) *Problem solved,* where each step is explained, always starting with a statement of what is being asked[8], then explicitly state the data already provided in the text of the problem

[8] At the time when the user studies were being conducted, we realised that when asked to solve a problem many children could not even determine what the problem was and what they were supposed to find.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

followed by the formula or arithmetic operation needed to obtain the required result, repeating this process as many times as needed. At the end, the result is highlighted in red and starts with a statement that refers to the initial one.

3.4 Typographical design

Our purpose for presenting the information is to make the dictionary as interesting, appealing and as dynamic as possible in such a way that children would enjoy consulting it while at the same time, it is consistent and coherent. This is important because:

a) we want to engage the children and make them curious as to continue reading entries in addition to what they were originally looking for.

b) children sometimes perceive dictionaries as "instruments of punishment" when they behave badly and must copy several entries (a reason to dislike dictionaries in general), so we want to show them that dictionaries can open the door to learning interesting things and can also be fascinating.

To achieve this, we defined the following set of typographical specifications:

- The dictionary will be published in print form and the online version will be available later.
- There is a page per entry in the print edition.
- Pages do not have a fixed layout.
- Colour edition.
- Each data category is presented in specific colour, for example, the term is always written in blue, the definition in red, additional information in green, solved problem in black, and so on.
- Use of the largest font size for the available space.
- Easy to read font type.
- Illustrations may be placed anywhere suitable on the page.
- There is an illustration of the abstract concept and whenever convenient, one or more real-life objects depicting the application of the concept.

4. Concluding remarks

The final stages of the preparation were severely delayed for several reasons, amongst which the difficulties to find a professional illustrator to accommodate the project in times of Covid-19. Virtual meetings were the new normal, but we had to learn how to communicate the ideas for each illustration without the closeness of face-to-face explanations.

On the other side, there were additional user studies we wanted to carry out to evaluate the final dictionary before going to press. However, because of the time lapse, many children who had taken part in the previous studies had already finished primary school. And all schools were closed for the best part of a year and a half.

So, the status of the print version of the dictionary is "soon to be published". The online version will come out a year after the printed version.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

With the publication of this dictionary, we hope to achieve the following:

a) help as many children as possible to understand basic maths concepts,

b) facilitate the establishment of links between related concepts in a way that each child may structure his or her maths knowledge,

c) establish a link between abstract concepts and their application to real life through well-chosen examples and vice-versa,

d) help children to learn the general procedure to approach a maths problem and solve it in a satisfactory way, by dividing it into partial steps,

e) have access to interesting and appealing information connected with the term children are consulting,

f) enhance maths knowledge in combination with the corresponding textbook,

g) hopefully, by understanding the concepts they learn, children will like the subject and therefore make the understanding of maths easier and more enjoyable.

References

Children's dictionaries

Bana, J./Marshall, L./Swan, P. (2006): The complete handbook of maths terms. Wexford.

Les grands et petits dictionnaires Dictionnaire Mathematique: Paris.
https://www.dictionnaires.com/mathematique/.

Frith, A./Lacey, M. (2008): See inside maths. Saffron Hill.

Gardner, K. (2005): Collins maths dictionary. London.

Gaspard, S. (2021): El Vocabulario Matemático Esencial.
https://www.superprof.mx/blog/diccionario-de-matematicas/.

Induráin, J. (2009): Diccionario Esencial Matemáticas, México City.

Jason Abdelnoor, R. E./Chandler, S. (2006): A maths dictionary. Cheltenham.

Le Robert Junior des Maths (2021): Mathématiques Illustrées de A à Z. Paris.

Litton, J./Flintham, T. (2014): Matemáticas Maravillosas. Mexico City.

Patilla, P. (2008): Oxford primary maths dictionary. Oxford.

Pierce, R. (2020): Disfruta las matemáticas: https://www.disfrutalasmatematicas.com/definiciones/.

Rich, G. (2002): Maths dictionary 11–14. London.

Turner, G. (2003): Developing numeracy. Primary maths dictionary. London.

General references

Barriga, R. (2022): Oct 2019. Los libros de texto gratuitos en su 60 aniversario. In: Otros Diálogos, no. 19, April–June 2022, El Colegio de México, México City: OTROS DIÁLOGOS | Los libros de texto gratuitos en su 60 aniversario (colmex.mx) (last access: 21-04-2022).

Cabré, M. T. (1993): La Terminología. Teoría, Metodología, Aplicaciones. Barcelona.

Cabré, M. T. (1999): La Terminología. Representación y Comunicación. Barcelona.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Cabrera-Diego, L. A./Sierra, G./Vivaldi, J./Pozzi, M. (2011). Using Wikipedia to validate term candidates for the Mexican Basic Scientific Vocabulary. In: First International Conference on Terminology, Languages, and Content Resources (LaRC 2011). Seoul, pp. 76–85.

Estopà, R. (2021): Los corpus infantiles elementos clave para la elaboración de diccionarios especializados escolares basados en el aprendizaje significativo botton-up y en el principio de adecuación. In XVII Simposio Iberoamericano de Terminología RITerm, October 26–29, 2021, Mexico City.

International Monetary Fund (2021): World Economic Outlook Database: https://www.imf.org/en/Publications/WEO/weo-database/2021/October/weo-report? (last access: 14-03-2022).

OECD 2019 (2019): PISA 2018 results. OECD, 3 December 2019. https://www.oecd.org/pisa/publications/pisa-2018-results.htm (last access: 14-03-2022).

Piaget, J. (1967): Logique et Connaissance Scientifique, Encyclopédie de la Pléiade. Paris.

Pozzi, M. (2016): Design of a corpus for mathematical knowledge transfer to 6-to-12-year-old children. In: Erdman Thomsen, E./Pareja-Lora, A./Nistrup Madsen, B. (eds.): Term bases and linguistic linked open data. Copenhagen, pp. 112–123.

Sinclair, J. (1995): Collins COBUILD English dictionary. London.

Vygotsky, L. S. (1978): Mind in society. The development of higher psychological processes. Cambridge, MA.

Vivaldi, J./Cabrera-Diego, L. A./Sierra, G./Pozzi, M. (2012): Using Wikipedia to validate the terminology found in a corpus of basic textbooks. In: LREC 2012, pp. 820–827.

Wild, K./Kilgarriff, A./Tugwell, D. (2013): The Oxford Children's Corpus: using a children's corpus in lexicography. In: International Journal of Lexicography 26 (2), June 2013, pp. 190–218.

Contact information

María Pozzi
El Colegio de México
pozzi@colmex.mx

Acknowledgements

I wish to thank the Consejo Nacional de Ciencia y Tecnología (CONACYT) for the generous funding of this research project: *El vocabulario básico científico de México. Un estudio de sus características, componentes y difusión* (No. CB-2013-220528-H).

I also wish to thank *Colegio Madrid*, *Colegio Oxford para niñas* y *Colegio Oxford para niños*, the three schools in Mexico City that allowed their students to participate in the user studies.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Stefan J. Schierholz/Monika Bielinska/
Maria José Domínguez Vázquez/Rufus H. Gouws/
Martina Nied Curcio

THE EMLex DICTIONARY OF LEXICOGRAPHY (EMLexDictoL)

Abstract The EMLex Dictionary of Lexicography (= EMLexDictoL) is a plurilingual subject field dictionary (in German, English, Afrikaans, Galician, Italian, Polish and Spanish) that contains the basic subject field terminology of lexicography and dictionary research, in which the dictionary article texts are presented in a sophisticated but comprehensible form. The articles are supplemented by a complex cross-referencing system and the current subject field literature of the respective national languages. Following the lemma position, the dictionary articles contain items regarding morphology, synonymy, the position of the definiens, additional explanations, the cross-reference position, the position for literature, the equivalent terms in the other six languages of the dictionary as well as the names of the authors.

Keywords Special field lexicography; multilingual dictionary; EMLex

1. Introduction

In lexicography and dictionary research the possibilities to find special field terminology of a reliable quality are restricted. Although the "Wörterbuch zur Lexikographie und Wörterbuchforschung/Dictionary of Lexicography and Dictionary Research" (2010–2020)[1] is a comprehensive terminologized und multilingual special field dictionary, its intended users are primarily experts in the subject field, lecturers, practical lexicographers and to a lesser extent students of lexicography, linguistics or translation studies. Other special field dictionaries of lexicography are limited with regard to their extent or to a single language.[2] Only a few of these special field dictionaries contain a didactic component directed at students of the subject field of lexicography whose first language is not the language of teaching used in their linguistic or lexicographic training. There is a lack of a didactically compiled plurilingual special field dictionary that could be of significant assistance to students with different mother-language backgrounds, and also to lecturers that have to teach in different languages. Although monolingual special field dictionaries of linguistics might be available for terminological questions from the special field of lexicography, they do not only display an insufficient lemma coverage regarding linguistics,[3] but omit the special field terminology of lexicography and dictionary research, or only treat it inadequately.[4]

The conception of the EMLex Dictionary of Lexicography is the initiative of Monika Bielinska (University of Silesia, Poland) and was then executed in collaboration with lecturers in the programme "European Master in Lexicography" that work in Germany, Italy, Spain and South Africa as dictionary researchers and lexicographers. It is a special field dictionary that

[1] Wiegand et al. (2010–2020).

[2] For example Swedenborg (1990); Yang/Xu (1992); de Sousa (1995); Bergenholtz et al. (1997); Hartmann/James (1998/2000); Burkhanov (1998/2010); Bielińska (2020).

[3] See Kreuder (2003) and Schierholz/Wiegand (2004) for the situation regarding German.

[4] For German Bußmann (2008) and Glück/Rödel (2016) among others should be mentioned.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

contains the basic special field vocabulary of lexicography and dictionary research, and the dictionary article texts are presented in both a sophisticated specialised and comprehensible way.

2. The dictionary type

The EMLex Dictionary of Lexicography is a plurilingual[5] didactic, terminologized special field dictionary for laypeople with a basic knowledge of linguistics, semi-experts and experts. It has a poly-alphabetical macrostructure and contains dictionary articles in German, English, Polish, Italian, Afrikaans, Spanish and Galician, presenting two languages (Galician and Polish) not included in the WLWF, where, besides German and English, the lemmata were translated into Afrikaans, Bulgarian, French, Italian, Portuguese, Russian (partially), Spanish and Hungarian.

EMLexDictoL has been conceptualized as an extension dictionary and should be produced as both an online and a printed dictionary. The complete dictionary consists of seven separate special field dictionaries in the above-mentioned languages. Hereby the lexicographic terminology is captured and treated in different languages and a contribution is made to the expansion and standardization of the most important lexicographic terms, so that the teaching of subject field communication and the dictionary culture can be promoted.

The article texts in EMLexDictoL are not translations but rather adaptations of the German version, so that examples, dictionary excerpts and illustrations as often as possible come from the specific language. Articles are supplemented by a complex cross-reference system and the relevant subject field literature from the specific national languages.

The conceptualization as extension dictionary pursues two distinct objectives: a) The complete dictionary with its seven separate special field dictionaries in the above-mentioned languages could have further languages added to extend its scope. b) The lemma candidate list can be extended if it is determined that the terms collected for the first phase are insufficient to present the basic vocabulary of lexicography and dictionary research.

3. Target users and functions of the EMLexDictoL

The genuine purpose of the EMLexDictoL is to be a reference work in the teaching of lexicography that can be used by students and lecturers. Therefore, the primary target users of the EMLexDictoL are students in the field of lexicography as well as students that need basic knowledge of lexicography in their studies or want to look it up. These students might often be studying linguistics but also within their training in another subject they might have to deal with lexicography or special field lexicography when having to plan or compile a special field dictionary or glossary in that study field. Due to the plurilingualism of EMLexDictoL its target user group is relatively big; not only because this specialised dictionary can be used in different countries but also because it can be employed in specialised translation situations. The EMLexDictoL should therefore support translation, text reception, text pro-

[5] Here the expression "plurilingual" is used as a subject field term which indicates that the complete special field dictionary "EMLexDictoL" consists of different partial subject field dictionaries with one of the seven languages allocated to each one, but individually remain multilingual because they contain translations of the lemmata in the other languages.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

duction as well as systematic specialised research, and in addition also supports the lexicographic process.

The complete structure of the EMLexDictoL has the function to support standardisation with a focus on the current basic vocabulary in different languages.

4. The dictionary basis and the lemma selection

The German version, of which 90 lemmata are currently being treated, will serve as basis. Each separate dictionary in the other languages will contain the equivalents of the German lemmata. The compilation of the first list of lemma candidates was done with the expertise of the editorial team but it is not yet complete.

The basis for the lemma selection is in the first instance the teaching experience in the international programme, the "European Master in Lexicography" (EMLex) that has English and German as languages of instruction. In this regard it reflects the central terms in lexicography and dictionary research that are required in the teaching, or those that students frequently use incorrectly. Secondly the basis for the lemma selection is the available introductory textbooks in the subject field,[6] but also the bilingual systematic introduction in the WLWF[7] and basic works that are used during the EMLex training.[8]

In the lemma selection care was taken to consider different fields of lexicography (for example dictionary typology, structures, use and compilation) in a balanced way.

The printed version does have some length restrictions, because the presumed 900 pages for the complete dictionary will allow only 150 pages for each separate dictionary, for as long as, as is planned in the first phase, only a single volume specialised dictionary is envisaged. Further development of the dictionary to include additional languages or more dictionary articles could lead to the production of multi-volume editions.

5. The text compound structure

The text compound structure[9] of the EMLexDictoL is the textual structure of the text constituents included in this dictionary. Because the EMLexDictoL consists of seven special field dictionaries it displays a primary and a secondary text compound structure.

The primary text compound structure consists of the primary front matter, seven special field dictionaries and the primary back matter. The primary front matter contains the front cover, the title pages in seven languages ("EMLex-Wörterbuch zur Lexikographie / EMLex Dictionary of Lexicography / EMLex Leksikografiewoordeboek / EMLex Diccionario de lexicografía / EMLex Dizionario di lessicografia / EMLex Słownik leksykografii / EMLex Diccionario de lexicografía"), the preface in seven languages and the table of contents in seven languages. Then the seven special field dictionaries follow, each with its secondary text

6 Engelberg/Lemnitzer (2009),

7 Wiegand et al. (2010, pp. 1–242).

8 See among others the HSK volumes.

9 For "text compound structure" the term "macrostructure" is also used. See among Engelberg/Lemnitzer (2001, pp. 116 ff.).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

compound structure, and then the primary back matter that contains the complete literature list, the complete list of dictionary titles, and the list of abbreviations.

The complete literature list contains all bibliographic references of all seven special field dictionaries. In this list of references the shortened forms given in the respective literature positions will appear with their full forms. The complete list of dictionary titles contains all dictionary titles occurring in the seven special field dictionaries. Titles of both general language dictionaries and special field dictionaries are given.

The complete list of abbreviations contains all abbreviations occurring in the seven special field dictionaries along with their respective full forms.

The secondary text compound structure occurs seven times and consists in each special field dictionary of the secondary front matter, the word list and the secondary back matter. The secondary front matter of the German special field dictionary that precedes the English special field dictionary "EMLex Dictionary of Lexicography" contains the title page ("EMLex-Wörterbuch zur Lexikographie"), the table of contents of the "EMLex-Wörterbuch zur Lexikographie" in German and the user guide, also in German. This is followed by the word list with German subject field terms as lemmata and then the secondary back matter, which contains the list of abbreviations in German. This is attached to the second secondary text compound structure.

6. Dictionary articles

Each single dictionary of EMLexDictoL contains both basic dictionary articles and cross-reference articles. The individual article positions of a basic dictionary article are now illustrated by means of the article of the monosemous term *Benutzer/Benutzerin*.[10]

All basic articles have a similar article structure. In the specific article positions the items are formulated as explicitly as possible and item texts are used extensively in order to avoid textual condensation as far as possible.

The "lemma position" is only populated by a single lemma. This could be a simplex (e. g., **definition, lemma**) or a compound (e. g., the German **Lemmaliste** (lemma list), **Textverbundstruktur (**text compound structure)). When a lemma designates a person (e. g., the German **Benutzer** (user)) the lemma with the gender suffix is given and separated by a slash from the first form in the lemma position (**Benutzer/Benutzerin**). The lemma sign could be a single word term or a multiword term (e. g., **bilingual dictionary**). For those terms that might have pronunciation problems an item giving the pronunciation is given.

The position for "Items regarding morphology" firstly contains an item regarding the use of the article and then – following a semicolon as structural indicator – the abbreviation "*Pl*" (for *plural*) in italics, the item regarding the number (for the lemma **definition** *definitions* are given). When the lemma with the gender suffix is given in the lemma position, the item regarding the use of articles contains the lemma with the gender suffix after a slash ("/") (*der Benutzer/die Benutzerin*) and in the item regarding the number the plural form is given (*Benutzer/Benutzerinnen*).

[10] See Figure 1.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

The position for synonyms can accommodate at the most three synonyms that are separated by semicolons and appear in alphabetical order. The position for synonyms is introduced by the "also" so that an identifying item for this position exists. For each item giving a synonym there only is one cross-reference article in which the synonym is indicated by the lemma so that the cross-reference is of a bidirectional type. A synonym relation can also be indicated for the long or short form of a term presented as lemma. In the example article in Figure 1 (see there) the long form *der Wörterbuchbenutzer/die Wörterbuchbenutzerin* (*dictionary user*) is given as synonym for the lemma **Benutzer/Benutzerin** (user).

Because EMLexDictoL is a special field dictionary the definitions of the terms in the lemma position are formulated to convey subject field constituting meaning knowledge to the users. This is never a sentence nor a sequence of sentences, but a phrase. Likewise, more than one sentence will not appear in the position for the definition and no abbreviations will be used.

By means of the article texts in the "Position for further explanations" the subject matter constituting meaning knowledge regarding the term should be deepened and expanded. All forms of images, e. g., scanned photographs, pictures of manuscripts, technical diagrams, tree diagrams, language maps, tables and dictionary articles may be used. Where complete dictionary articles or fragments are inserted, the typography in the dictionary articles should resemble the original as close as possible. Each figure or table has a caption or heading.

Metalexicographic references to the object language in the running text are highlighted by means of *italics* (e. g., "the term *Lemma sign* is […]"). Italics are not used to give special emphasis to a term. Quotation marks are only used for citations.

Dictionary articles inserted as examples are treated as figures and have their own caption in which the source of the article is also given as precisely as possible.

In the running text an arrow that points upwards ("↑") is used when reference is made to another term that occurs in EMLexDictoL as lemma ("↑lexikalisches Informationssystem" in Fig. 1).

In the additional explanations the effort is still made to ensure, by means of didactic-pedagogical text formulation, a high degree of comprehension for those people who receive training. Potential mistakes are also indicated to assist with the prevention of mistakes, cf. the section in the article of the lemma **Benutzer/Benutzerin** (Fig. 1)*:*

> In German specialised literature the form 'Nutzer' is often used. However, this should be avoided because a 'Nutzer' is a person that uses or applies a specific offer. In the field of dictionary research, the term 'Benutzer' is correct. 'Nutzer' should also not be confused with '↑addressee.' Addressees are persons at whom a dictionary is directed and for whom the dictionary was compiled because of a specific interest. They are not necessarily the users but potential users.

The cross-reference position is introduced by an arrow that points upwards ("↑"). It is followed by terms, in alphabetical order, that are included in EMLexDictoL as lemmata of articles with a basic article structure. Here terms are given that do not appear in the running text of the additional explanation and that contribute to the expansion of knowledge of the lemma sign.

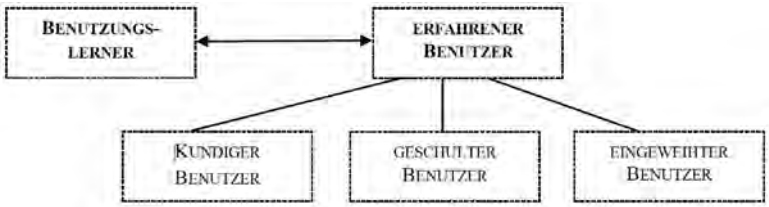
It is followed by two positions for literature: the position for dictionaries mentioned in the article text or as source of the examples being used starts at the beginning of the line with

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

the identifying indicator realised by the iconic sign "📖" that presents an open book (in Fig. 1 this position is empty). Only the dictionary title is given, followed by the date of publication. When more items occur in this position for literature they are ordered alphabetically.

The second position for literature contains subject field literature used in the composition of the specific article as well as recommended additional subject field literature. It is introduced by the iconic sign with two open books ("📖📖"). An item giving a reference to literature is entered as a shortened item (in Fig. 1 e.g. "Tarp (2009)") that contains the full surname of the author, in round brackets the year of publication followed by a colon and the page number(s) referred to in the article text. In both the literature positions shortened items are always used in order to limit the size of the dictionary article. At the end of EMLexDictoL an alphabetical list is given with the shortened items and the full items giving the literature of all items giving literature in all seven special field dictionaries.

The position for equivalent terms is introduced at the beginning of the line with an upright diamond ("◇") that serves as identifying indicator for this position. The equivalent position is populated by the equivalents of the other six languages of the subject field dictionary. A maximum of three equivalents are allowed for each language and they are presented in alphabetical order and separated by a semicolon. The name or names of the author or authors appears at the end of the dictionary article text.

lemma position	Benutzer/Benutzerin
position for morphology	der Benutzer/die Benutzerin; *Pl.* Benutzer/Benutzerinnen
position for synonymy	auch: der Wörterbuchbenutzer/die Wörterbuchbenutzerin
definition	Person, die ein Wörterbuch oder ein lexikalisches Informationssystem zum Nachschlagen, Überprüfen oder Üben verwendet.
position for further explanations	Es können verschiedene Typen von Benutzern unterschieden werden: In Bezug auf den Grad der Kompetenz kann man einerseits den Benutzungslerner (Person, die lernt, ein Wörterbuch zu benutzen und es in der Praxis übt) und andererseits den kundigen Benutzer (Person, die ein bestimmtes Wörterbuch gut kennt), den geschulten Benutzer (Person, die potentiell ein Wörterbuch adäquat benutzen kann) und den eingeweihten Benutzer (Person, die ein Wörterbuch praktisch benutzen kann und auch das metalexikographische Wissen dazu hat) unterscheiden (Wiegand 1998, S. 505–508). Der kundige, geschulte und eingeweihte Benutzer gehören zur Gruppe der erfahrenen Benutzer und sind Personen, die Erfahrung in der Wörterbuchbenutzung haben und aufgrund ihrer Kenntnisse, ihres Wissens sowie ihrer praktischen Erfahrung als potentielle Benutzer ein Wörterbuch adäquat benutzen können.  **Abb. 1:** Typen von erfahrenen Benutzern

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

position for further explanations	Außerdem wird die Unterscheidung von erfolgreicher vs. erfolgloser Benutzer, muttersprachlicher vs. fremdsprachlicher Benutzer, sowie Laienbenutzer vs. wissenschaftlicher Benutzer vorgeschlagen (Wiegand 1998, S. 509–510). Am besten erforscht ist der fremdsprachliche Benutzer (Wiegand 1998, S. 264). Nimmt man die Benutzungshandlung und die Benutzungssituation in den Fokus, so ist es möglich, zwischen potentiellem Benutzer (Person, die vielleicht ein Wörterbuch benutzen wird), Benutzer in-actu (Person, die gerade dabei ist, ein Wörterbuch zu benutzen), Benutzer ex-actu (Person, die ein Wörterbuch benutzen kann, es aber im Moment nicht tut) und Benutzer post-actu (Person, die an einem Text arbeitet und dafür zuvor ein Wörterbuch benutzt hat) zu differenzieren (Wiegand 1998, S. 501). «Wörterbücher sind Gebrauchsgegenstände, die benutzt werden» und «[j]eder, der ein Wörterbuch benutzt, handelt» (Wiegand 1998, S. 262). Der Benutzer steht aufgrund dieses handlungsorientierten Ansatzes im Zentrum der Wörterbuchbenutzungsforschung. Es ist wichtig, dass diese Person bereits Erfahrung im Umgang mit einem Wörterbuch oder einem ↑lexikalischen Informationssystem hat und dass die Benutzerstudie sich auf den Benutzer in-actu konzentriert. Eine Erforschung des Benutzers ist wichtig, damit sich Lexikographen und die lexikographische Produktion gezielt an den Benutzungsbedürfnissen, den konkreten Benutzungshandlungen und ihrer Benutzungssituation und ihrem Benutzungskontext, sowie der Benutzungskompetenz und an eventuellen Benutzerfehlern orientieren und ihre Wörterbücher bzw. lexikalischen Informationssysteme benutzerfreundlicher gestalten können. Manchmal wird in der deutschen Fachliteratur auch von Nutzer gesprochen. Dies sollte jedoch vermieden werden, denn ein Nutzer ist eine Person, die ein bestimmtes Angebot nutzt oder anwendet. Im Bereich der Wörterbuchforschung sind die Termini Benutzer und Benutzerin korrekt. Benutzer sollte außerdem nicht mit ↑Adressat verwechselt werden. Adressaten sind Personen, an die sich ein Wörterbuch wendet und für die das Wörterbuch aufgrund eines spezifischen Interesses erstellt wurde. Sie sind nicht unbedingt die Benutzer, sondern potentielle Benutzer.
reference position	↑Nachschlagewerk, Wörterbuch, Wörterbuchbenutzung, Wörterbuchbenutzungsforschung
position for dictionaries	🕮
position for literature	🕮🕮 Tarp (2009); Wiegand (1998, S. 262, 264, 501, 505–510); WLWF-1/2010, S. 675–678, 681, 685–686, 688; WLWF-2/2017, S. 146, 206–207, 408, 571; WLWF-3/2020, S. 183.
position for equivalents	◇ **afr** *gebruiker* **en** *user* **es** *usuario* **gal** *usuario* **it** *utente* **pl** *użytkownik*
position of the author's name	Martina Nied Curcio

Fig. 1: Example of a dictionary article

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

7. The mediostructure

Mediostructural elements often occur in EMLexDictoL because conceptual scientific knowledge is coherent knowledge that cannot be portrayed by the alphabetical macrostructure of the subject field dictionary. For the active user the knowledge coherence that has been neglected by the dictionary-internal data distribution can be reconstructed by the observance of the cross-reference offer. "Reconstructed" means that it must have been constructed before. For this purpose, a complex mediostructure was devised in the conception of EMLexDictoL. It has always been populated with contents by the authors of the dictionary article texts, by entering the relevant content-related and additional connections on the basis of the different cross-reference type options. Thereby a mediostructural selection will take place because not all cross-references but only the function-relevant cross-references are given, and because each of the single dictionary functions will require a different mediostructural selection. This means that EMLexDictoL in which the text reception function precedes the specialised information function, can be used successfully.[11]

By means of the positional structuring of the articles of the subject field dictionary the text reception function can already be achieved when an active user, due to a knowledge gap regarding a searched term, finds the lemma in EMLexDictoL and reads and understands the specialised definition, so that the knowledge gap can be closed. The active user can deepen his knowledge when he also reads and understands the additional explanation in the respective dictionary article. He can recognise content-related terms when following the cross-references in the additional explanation or the cross-references in the cross-reference position, so that the consultation executed to fill the knowledge gap becomes a specialised enquiry.

In addition, the active user is supported by the article-internal cross-reference to the literature position that guides him to specialised literature found in abbreviated form in the literature position. The items giving the shortened forms are unidirectionally connected to the complete reference list in which the solution to the shortened forms is given so that all the items giving the literature can be found there.

In the same way, the abbreviations of dictionary titles and author names in the article text or the literature position are unidirectionally connected to the complete list of dictionary titles and of author names respectively.

The equivalent terms ordered according to languages in the equivalent position are the carriers of cross-reference addresses, guiding the user from the article to the other seven subject field dictionaries in EMLexDictoL.There the given terms are included as lemmata of a full subject field dictionary article or cross-reference article in the target language.

8. The actual standing of the work on the EMLexDictoL

The project is still in its initial phase. After a comprehensive planning the first test articles were compiled in 2021. In the first phase about 90 lemmata, selected with the expertise of the team of editors, were treated. The work should be concluded during 2023. The online version is in its preparatory phase and will be produced parallel to the printed version. The authors are aware of the recent research situation where in most cases only online dictionaries are of interest for lexicography and dictionary research. But being involved in the

[11] Regarding the mediostructure, see among others Wiegand (2002 [2003], pp. 216–222).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

recent education process of young lexicographers in the EMLex programme, being confronted with the problems young semi experts in lexicography experience when using metalexicographical terminology, and being familiar with education problems in third world countries, we regard the production of print dictionaries and the further theoretical development of structures for print dictionaries as important as the development of online dictionaries which are available on computers, tablets and smartphones.

References

Bergenholtz, H./Cantell, I./Vatvedt Fjeld, R./Gundersen, D./Jónson, J. H./Svensén, B. (1997): Nordisk Leksikografisk ordbok. Oslo.

Bielińska, M. (ed. 2020): Leksykografia. Słownik specjalistyczny. Kraków.

Burkhanov, I. (1998/2010): Lexicography. A dictionary of basic terminology. Rzeszów.

Bußmann, H. (ed.) (2008): Lexikon der Sprachwissenschaft. 4., durchges. und bibliogr. erg. Aufl. Stuttgart.

Engelberg, S./Lemnitzer, L. (2009): Lexikographie und Wörterbuchbenutzung. 4., überarb. und erw. Aufl. Tübingen.

Glück, H./Rödel, M. (eds.) (2016): Metzler Lexikon Sprache. 5., aktual. und bearb. Aufl. Stuttgart/Weimar.

Hartmann, R. R. K./James, G. (1998/2000): Dictionary of lexicography. London/New York.

HSK 5.1–5.3 (1989–1991) = Hausmann, F. J./Reichmann, O./Wiegand, H. E./Zgusta, L. (eds.):Wörterbücher. Ein internationales Handbuch zur Lexikographie. (= Handbücher zur Sprach- und Kommunikationswissenschaft (HSK) 5.1–5.3). Berlin/New York.

HSK 5.4 (2013) = Heid, U./Gouws, R. H./Schweickard, W./Wiegand, H. E. (eds.): Dictionaries. An international encyclopedia of lexicography. Supplementary volume: Recent developments with special focus on computational lexicography. (= Handbücher zur Sprach- und Kommunikationswissenschaft (HSK) 5.4). Berlin/New York.

HSK 14.2 (1999) = Hoffmann, L./Kalverkämper, H./Wiegand, H. E. (eds.): Fachsprachen. Ein internationales Handbuch zur Fachsprachenforschung und Terminologiewissenschaft. (= Handbücher zur Sprach- und Kommunikationswissenschaft (HSK) 14.2). Berlin/New York.

Kreuder, H.-D. (2003): Metasprachliche Lexikographie. Untersuchungen zur Kodifizierung der linguistischen Terminologie (= Lexicographica. Series Maior 114). Tübingen.

Schierholz, S. J./Wiegand, H. E. (2004): Die Wörterbücher zur Sprach- und Kommunikationswissenschaft (WSK). In: Lexicographica 20, pp. 164–264.

Sousa, J. M. de (1995): Diccionario de lexicografía práctica. Barcelona.

Swedenborg, L. (1990): En liten ordbok om ordböcker. Stockholm.

Tarp, S. (2009): Reflections on lexicographical user sesearch. In: Lexikos 19, pp. 275–296.

Wiegand, H. E. (2003): Altes und Neues zur Mediostruktur in Printwörterbüchern. In: Lexicographica 18/2002, pp. 168–252.

Wiegand, H. E./Beißwenger, M./Gouws, R. H./Kammerer, M./Mann, M./Storrer, A./Wolski, W. (2010–2020): Wörterbuch zur Lexikographie und Wörterbuchforschung. Bd. 1–5. Berlin/Boston.

Yang, Z./Xu, Q. (1992): Cishuxue cidian. Shanghai.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Contact information

Stefan J. Schierholz
Friedrich-Alexander-Universität Erlangen
Stefan.Schierholz@fau.de

Monika Bielinska
Uniwersytet Śląski
monika.bielinska@us.edu.pl

Maria José Domínguez Vázquez
Universidade Santiago de Compostela
majo.dominguez@usc.es

Rufus H. Gouws
Stellenbosch University
rhg@sun.ac.za

Martina Nied Curcio
Università degli Studi Roma Tre
martina.nied@uniroma3.it

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Historical Lexicography: German

Volker Harm

WORTGESCHICHTE DIGITAL: A HISTORICAL DICTIONARY OF NEW HIGH GERMAN

Abstract *Wortgeschichte digital* ('digital word history') is a new historical dictionary of New High German, the most recent period of German reaching from approximately 1600 AD up to the present. By contrast to many historical dictionaries, *Wortgeschichte digital* has a narrated text – a "word history" – at the core of its entries. The motivation for choosing this format rather than traditional microstructures is briefly outlined. Special emphasis it put on the way these word histories interact with other components of the dictionary, notably with the quotation section. As *Wortgeschichte digital* is an online-only project, visualizations play an important role for the design of the dictionary. Two examples are presented: first, the "quotation navigator" which is relevant for the microstructure of the entries, and, second, a timeline ("Zeitstrahl") which is part of the macrostructure as it gives access to the lemma inventory from a diachronic point of view.

Keywords Historical lexicography; word history; quotations; visualizations

1. Introduction: How to tell a word's history

When historical lexicography started during the 19th century, words were very often imagined as living beings whose biography had to be told in a dictionary – "every word should be made to tell its own history", as Herbert Coleridge has put it in a famous statement which sketched the new lexicographical enterprise that later became famous as the *Oxford English Dictionary* (cited from Mugglestone 2016, p. 556). Thus, dictionaries like the OED or the *Deutsches Wörterbuch* by the brothers Grimm are not only designed for looking up words, but also for reading and studying a word's history. Jacob Grimm, in his preface to the first volume of his *Deutsches Wörterbuch*, has made this quite clear: "The dictionary could be read at home, with pleasure and sometimes with devotion, too" ([1]DWB 1, p. xiii).[1]

While working on the last fascicles of the second edition of the *Deutsches Wörterbuch* in 2016, we started thinking about new ways of doing historical lexicography in the internet age. We always bore those quotations in mind: Our dictionary should tell the history of words much in the sense of Coleridge and Grimm, and a user should be given the opportunity to learn about a word's history by reading in our dictionary. However, to really meet these requirements traditional entry structures seemed insufficient to us. In our view, the major shortcomings in these structures are their opacity, the isolated treatment of individual words due to the alphabetical order, and, most importantly, their lack of a genuinely historical approach. These traditional structures are opaque as many users do not know how to extract useful information out of a list of definitions and quotations which are amassed in such an entry, often without comments or explanations. These structures prove insufficient in many cases as traditional dictionaries do not account for the fact that many developments can only be understood when words are considered within the context of their paradigmatic relations – an insight well known to linguistics ever since Trier (1931). And after all, these structures do not offer a historical approach as they provide nothing more than a

[1] Translation by the author (original: "so könnte das wörterbuch zum hausbedarf, und mit verlangen, oft mit andacht gelesen werden").

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

chronological arrangement of word meanings and quotations. But, as every historian knows, history is much more than chronology: instead of cataloguing word senses by their earliest attestation, one has to explain contexts, to show connections, and to sketch developments. Thus, the so-called "historical principle" of lexicography (Considine 2016, p. 163) has to be reinterpreted and adapted to our present needs and research questions.

2. *Wortgeschichte digital*: An online-only dictionary with narrated entries

In order to be clear about the aim of the new dictionary the project was termed *Wortgeschichte digital* ("digital word history", WGd), which is a nod of course to the great lexicographers of the 19th century whose intention was to provide word histories in their dictionaries. The inclusion of the adjective *digital* is not just a concession to fashion. Rather, the attribute points to an important hallmark of the project: It is, at least as far as I can see, the first online-only dictionary in the field of German language history. There is no retro-digitized dictionary from which the work starts, nor is there any print edition as a supplementary product. The decision to work digitally from the very beginnings is fundamental and has far-reaching consequences for the design of our dictionary. I will come back to this later.

Next to its digital nativity, the most outstanding feature of WGd is its style of presentation which at first sight, is not digital at all, but deeply rooted in the era of printing: Rather than using traditional microstructures consisting of definitions, usage labels, and quotations we opted for narrative entries, i. e. for articles written as continuous texts. In this respect, WGd has more in common with, e. g., Alain Rey's *Dictionnaire historique de la langue française* (DHLF) or Raymond Williams's *Key Words* than with the OED or the DWB. There are mainly two closely connected reasons for this decision: the one has to do with historical lexicology as the object of the dictionary, and the other with the dictionary's users.

The first motivation for turning to a narrative style is that the realities of word history are "complex and often messy" (Durkin 2016, p. 252). There is a number of issues which the lexicographer has to cope with: The word's origin and development has to be determined against the background of a sometimes scarce attestation, the successive emergence of new meanings out of older ones has to be reconstructed while accounting for the whole scale of the word's diatopic, diastratic, and diaphasic variation, and more than often the specific historical setting of a word usage has to be determined in order to properly contextualize the semantic changes. And after all, the mechanisms that brought about the lexical innovations – mostly metaphors and metonymies – have to be identified, which is not always a straightforward task. All this makes it difficult to deal with word histories within the constrictions of traditional entry structures. Those structures tend to prefer clear-cut distinctions between senses, each illustrated by well-fitting historical quotations. But when it comes to history, transitions are more relevant than well-established usages, and rather than the 'good' quotations it is the ambiguous attestations of a word which very often play a crucial role as bridging contexts for the rise of new meanings. In sum, the high flexibility which is necessary for dealing with the vagueness of many historical developments and for really living up to the complexities of historical lexicology can best be achieved in a continuous text.

As indicated above, there is another reason for turning to narrative entries. This style of representation is not only appropriate to the matter of word history itself, it is also more

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

easily comprehensible for the reader. The enumeration of senses, often forced into intricate semantic hierarchies, and large sections with quotations are unsatisfying for many users, especially if they are not used to reading historical texts. If readers expect answers to their questions into words and their developments, these answers should be spelled out as clearly as possible rather than to be hidden in the quotation section or in a labyrinthine entry structure. If dictionaries do not meet these demands, people interested in language will turn to other, less reliable, resources which can easily be found in the vastness of the internet.

Next to the decision to publish an online-only narrated dictionary, there is another feature which sets WGd apart from current historical dictionaries: its onomasiological approach. Onomasiology is relevant for the project design both with respect to the overall lemma selection and the individual entries. As to the selection of lemmata, WGd starts from conceptual domains such as, e. g., society, politics, economy, technology, communication, traffic, and everyday culture, thereby preferring those domains of vocabulary which underwent significant change from 1600 to the present. These domains are dealt with in individual work periods of several years. At present, WGd is dealing with the large area of *politics and society* for which five years are scheduled. The onomasiological approach has the advantage of offering a more economical way of describing the vocabulary, because instead of arbitrarily shifting from topic to topic following the contingencies of the alphabet, the lexicographers can focus on selected domains on which they work for a certain time. On the side of the user, this has the advantage of a far more coherent picture of the dictionary entries which are more or less from the same mould (for more detail on the onomasiological approach see Harm 2022, p. 175–177).

Onomasiology is also relevant on the level of the individual entry, although to a lesser extent. Here, next to the standard entries which are mostly dedicated to single words, one can also find synoptical entries describing word fields. One example is the field of terms for the upper classes containing lemmata such as *Beaumonde*, *die oberen Zehntausend*, or *Jetset*. The summarizing entry accounts for the interdependencies between the words and their developments and gives an overall picture of the ways a certain domain has been lexicalized over time.

3. Word history and quotation evidence in WGd

When it comes to telling word history in the sense of Coleridge and to compiling a dictionary as a "Lesebuch" in the vein of Jacob Grimm, the advantages of a narrated dictionary text are obvious. But there is a significant disadvantage of this dictionary type as the example of Rey's and Williams's dictionaries shows, namely the absence of quotation evidence. Whereas quotations are certainly not of primary interest for the ordinary reader, they play a crucial role for historical lexicography as a branch of philology. As far as dictionaries have scientific aspirations they cannot go without quotation evidence: quotations from historical sources are indispensable not only because they illustrate the word meanings in question, but also, and most importantly, because they provide the means for critically checking the lexicographer's hypotheses.

Most obviously, it is difficult to integrate a larger amount of quotations (including their bibliographic references) into a continuous text. An easy way to solve this problem would be to drastically shorten the quotations so as not to interrupt the text flow. But quotations which are too short are of little use because they miss their fundamental illustrative func-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

tion. In order to provide a solution to this issue, the DHLF, e. g., inserts merely a date into the text as an indication that there is an attestation of the word or word usage in question somewhere in the dictionary archives. In this case, the reader has to blindly trust the information he is given. *Trübners Deutsches Wörterbuch*, another prominent example of narrative lexicography, tried to solve the problem in a different manner: Selected quotations are given in a footnote appendix to each entry. Hereby, the reader gets access to at least some of the evidence the entry is based on, but the amount and quality of the information provided this way remains far too scarce altogether. Thus, in a printed dictionary, writing a continuous text and presenting the quotation evidence in a satisfactory manner is nearly impossible.

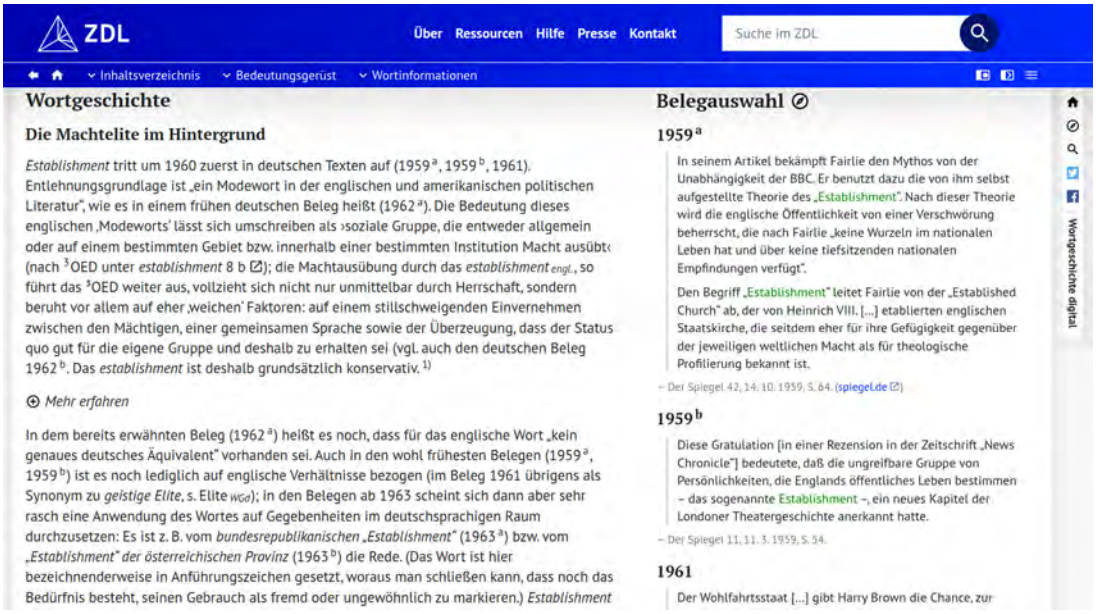


Fig. 1: Word history and quotation selection

It is only in a digital dictionary where both demands – presenting a readable text while not omitting the quotations – can be met. Therefore, WGd has developed a solution termed "movable quotation section" (see Fig. 1). The basic structure of this feature is as follows: the text and the quotation section are presented alongside on the screen, whereby the text (as the most interesting component) occupies the largest part and the quotation section is given on the right side. In order to show the relationship between text and quotations both are mutually connected by links (symbolized by the arrows in Fig. 1). In the text, the quotations are indicated by the year of publication of the source; by clicking on the date, the relevant quotation will be highlighted and scrolled automatically into view. The link is working the other way round too: Clicking on the quotation (i. e. the year index) will lead the reader to the place in the text where the quotation is dealt with. Whenever possible, the quotations themselves are connected with their source in the corpus[2] or on the internet by an external link.

[2] WGd mainly relies on current reference corpora of German such as Deutsches Textarchiv, a large text collection consisting of about 1500 sources from 1600 to 1900, and the DWDS-Korpus ranging from about 1900 up to the present, cf. https://www.deutschestextarchiv.de and https://www.dwds.de. Next to these corpora which are the standard references for the project, a number of other resources is used: library collections such as https://anno.onb.ac.at/, e. g., or the corpora provided by the *Leibniz-Institut für Deutsche Sprache* in Mannheim (IDS). – As larger internet-corpora like DTA and DWDS

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

By the joint presentation of both the word history and the quotation evidence it is based on at least two different goals can be reached. First, a broader audience can be addressed as both experts and laymen can find what they want: The professional reader has all the material at his disposal in order to critically examine the entries and to draw his own conclusions, and non-professionals will not be disturbed or distracted by information which is of no use to him unless he wishes to go into further detail. Second, interlinking word history and quotations has positive consequences for the lexicographical structure itself. In traditional dictionaries such as DWB, it is not always clear what lexicographical information a quotation is meant to illustrate or corroborate. In the mode of presentation used in WGd, however, the relationship is unambiguous since quotation and dictionary description are always closely tied together.

4. The microstructure of WGd entries

The ideal users of WGd are those who have a basic interest in and perhaps some knowledge of word history and who are willing to take the time to read such an entry. But this is not the only audience we have in mind. As normally dictionaries are the only interface between linguistics and a broader audience, we see it as our duty to address a wide array of readers. Therefore, the dictionary has to accommodate its structure to an audience consisting not only of experts. In order to make the entries more easily accessible, it is necessary to provide different modes of reading. Next to a close reading of the whole entry, the dictionary has to offer valuable information also for those who just want to get a short answer to a very specific question and to those who just like to browse through the dictionary.

In order to enable users to browse through the entries and to focus on individual sections of the entry, a special introductory component has been created which is located on the top of each entry: the so-called "orientation section" ("Orientierungsbereich"). As shown in Figure 2, this section includes:

a) a short abstract of the word history ("in short" / "Kurz gefasst")
b) a table of contents ("Inhaltsverzeichnis")
c) an overview of all the word meanings relevant for the entry ("Bedeutungsgerüst" / "relevant senses").
d) a section with additional information on the word ("Wortinformationen" with sense relations, collocations, word formations)

The components b) to d) are linked to the relevant heading or passage in the text. So, the user has direct access to the piece of information he might be interested in. This way, a selective reading of the entries is possible.

already exist there is no need for WGd to build up an extra text collection for the project. New digital sources which are of special interest to WGd can be implemented by the collaboration with the *Zentrum für digitale Lexikographie der deutschen Sprache* (ZDL), the overarching institution the project is part of.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

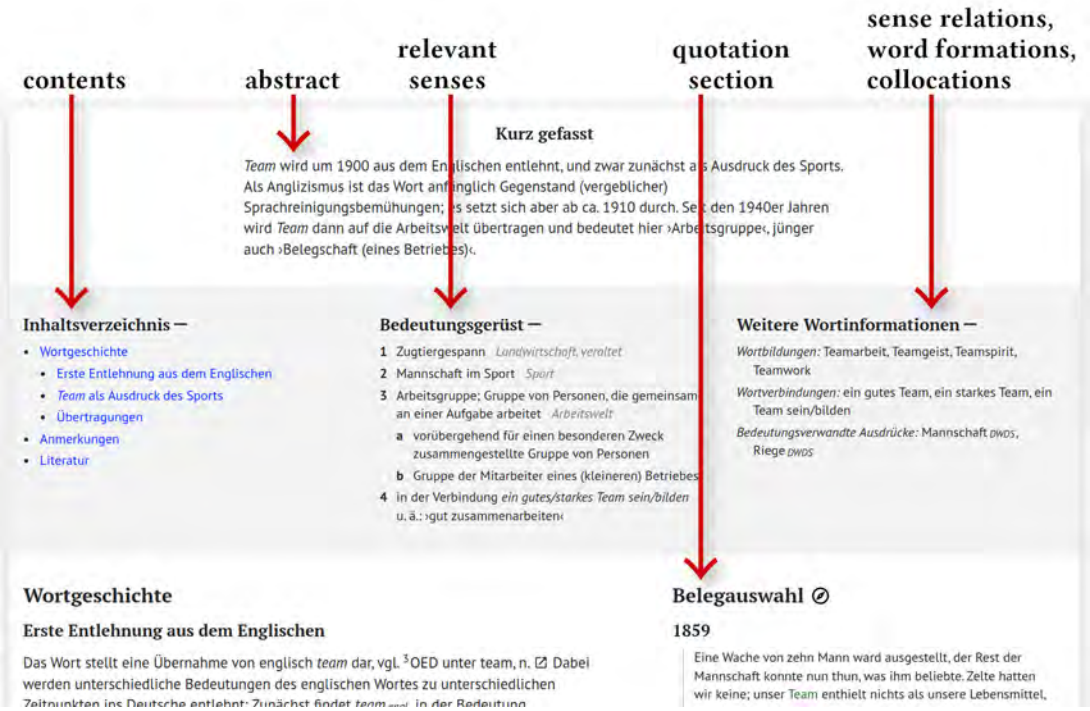


Fig. 2: Orientational section ("Orientierungsbereich")

As already mentioned, the quotation section normally includes a large number of quotations. In order to give an overview and to make this section more informative, a special feature has been introduced, the so-called "quotation navigator" ("Belegnavigator"; see Fig. 3.). This visualization can be opened by clicking a special icon. It is a timeline along which the quotations selected for the entry are listed as bullet points. These bullets are clickable, so that the reader can navigate to a quotation which he might find interesting: If one is interested in the 18th century, e. g., or in neologisms of the seventies, the bullets provide direct access to quotations from that period, and since the quotations themselves are interlinked with the passage where they are dealt with, the reader can also use this tool for a selective reading of the word history. The distribution of the bullets (i. e. the quotations they symbolize) can also give insights into historical facts which are not necessarily highlighted in the entry. Most notably, the hotspots of a word history, the periods which are critical for a word's semantic development may become visible by the distribution of the bullets. Figure 3, e. g., tells us that for the word *Establishment*, a loan from English, the first decade of its attestation (1959 to 1970) appears to be the most dynamic whereas since then no significant semantic innovations seem to have happened, at least according to the lexicographer who wrote the entry and selected the quotations. The suggestions provided by the quotation timeline might not always prove true, but, in any case, the tool can be seen as an invitation to the reader to deal with the content of the entry, to follow his own interests and even to draw his own conclusions. Most crucially, by the "quotation navigator" a playful component is introduced into the dictionary which provides alternative access to historical semantics.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Fig. 3: Quotation navigator

5. Macrostructures: the WGd timeline for lemma selection

Since in online dictionaries search windows offer very easy ways to look up words, macrostructures seem no longer necessary: A user who has just to type in what he wants to know does not need a structure leading him through the dictionary. Whereas it goes by itself that every online dictionary has to be searchable via a window, WGd aims at more: It seeks to arouse people's interest in things which they were not interested in or did not even know before. In order to reach this goal, it does not suffice to adopt the usual ways of public relations (notably Twitter and YouTube). The dictionary itself has to provide attractive tools enabling users to make unexpected discoveries. For this reason, WGd has a homepage including not only a search window and multifaceted filters, but also a complete lemma list with three optional arrangements:

a) by alphabet
b) by so-called "cross-reference clusters" which mirror the (mainly semantic) relations between the lemmata
c) a timeline ("Zeitstrahl")

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

As WGd is a historical dictionary the timeline is the most important tool for retrieving information. Therefore, it will be briefly illustrated here (for a detailed account of the cross-reference cluster see Dorn in this volume). If the user opts for the timeline view by clicking the relevant icon, all lemmata will be presented in chronological order according to their oldest attestation in the dictionary (which, by the way, is not necessarily identical to the earliest attestation in German altogether, because WGd is a dictionary covering New High German, the most recent language period reaching from approximately 1600 to the present). An extract of the timeline is given in Figure 4.

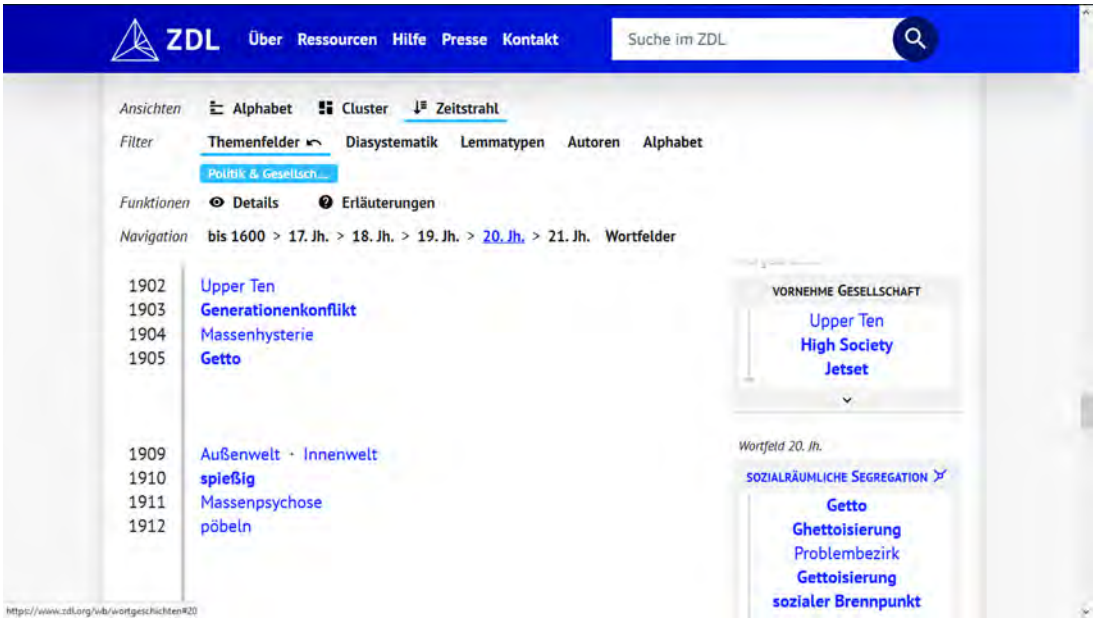


Fig. 4: WGd timeline ("Zeitstrahl")

On the left-hand part of the screen, the lemmata are listed in chronological order (the bold ones are so-called "Hauptlemmata"/"principal lemmata", which have an entry of their own, whereas the other lemmata are treated together with principal lemmata in a common entry). The chronological presentation offers the possibility of browsing through the lemma inventory following a personal interest in certain periods: Which words came about in the year of my birth? Which words emerged during the Second World War? It lies in the nature of such questions that they do not always have meaningful answers. Such a timeline, though, gives the reader the opportunity to explore the vocabulary on his own and to detect unexpected combinations and connections. However this may be judged, browsing through a timeline is certainly more informative than going through the alphabet.

As mentioned in section 4, WGd includes synoptical entries in addition to the standard entries. The lemmata dealt with in these entries can be found in the boxes on the right side (see again Fig. 4). By ticking the arrow below, the window opens and all relevant lemmata are displayed in chronological order (those in blue are from the relevant century which is selected). The lemma overview presented in these boxes, as simple as it is, can already reveal interesting facts about the evolution of the word field in question. In the case of the field "upper class terms" in Figure 5, e.g., it is obvious at first glance that the older words are mostly loans from French, whereas the words attested since the 20th century are of English origin. What is most important about this visualization is that every lemma appearing in the

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

timeline is linked with the respective entry. So, the reader is invited to directly go to the entry and read the text.

The lemma timeline as well as the quotation navigator explained in the previous section are created automatically. Nevertheless, these visualizations always rely on individual decisions made by lexicographers on the basis of their interpretation of the data. By integrating a larger number of these individual interpretations into a whole picture new perspectives and constellations become visible which reach beyond the single entries.

Fig. 5: Word field "upper class" in the timeline presentation

6. Conclusion

Wortgeschichte digital represents a significant departure from previous historical dictionaries as it has opted for "word histories", i.e. for a narrative format of representation, and for an onomasiological approach to lemma selection. Being an online-only dictionary from scratch, it makes extensive use of automatized visualizations which are complementary to the word histories. These visualizations offer new ways of accessing lexicographical information to the reader, who can more easily browse both through an individual entry and the dictionary as a whole. As we hope, these tools as well as the narrative entry format will make word history more attractive to the broad audience interested in words and language. But as important as digital working methods and visualizations are for the project, at its core, *Wortgeschichte digital* is a philological enterprise. The use of online corpora and digital tools does not change the fact that the interpretation of historical data is the cornerstone of the work on the project. In this respect, dictionary making has remained almost the same from the days of Jacob Grimm up to now: The aim of historical lexicography is still to write biographies of words, and this still needs highly skilled biographers in the first place.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

References

Considine, J. (2016): Historical dictionaries: history and development, current issues. In: Durkin, Ph. (ed.): The Oxford handbook of lexicography. Oxford, pp. 136–175.

DHLF: Rey, A.: Dictionnaire historique de la langue française. 1992, vols. 1–2. Paris.

Durkin, Ph. (2016): Etymology, word history, and the grouping and division of material in historical dictionaries. In: Durkin, Ph. (ed.): The Oxford handbook of lexicography. Oxford, pp. 236–252.

[1]DWB: Grimm, J./Grimm, W.: Deutsches Wörterbuch, vols. 1–16 [1–32]. Leipzig (1854–1971). https://www.woerterbuchnetz.de/DWB1. (last access: 25-03-2022)

[2]DWB (1960–2018): Grimm, J./Grimm, W. Deutsches Wörterbuch. Neubearbeitung, vols. 1–9. Leipzig/Stuttgart. https://www.dwds.de/d/wb-2dwb, https://www.woerterbuchnetz.de/DWB2. (last access: 25-03-2022).

Harm, Volker (2021): Beyond the "Grimm": German historical lexicography after *Deutsches Wörterbuch*. In: Van de Velde, H./Dolezal, F. (eds.): Broadening perspectives in the history of dictionaries and word studies. Newcastle upon Tyne, p. 117–134.

Harm, Volker (2022): *Wortgeschichte digital*. Ein neues Wörterbuch zur Geschichte des neuhochdeutschen Wortschatzes. In: Diehl, G./Harm, V. (eds.): Historische Lexikographie des Deutschen. Perspektiven eines Forschungsfeldes im digitalen Zeitalter. (= Lexicographica. Series Maior 161). Berlin/Boston, p. 173–191.

Mugglestone, L. (2016): Description and prescription in dictionaries. In: Durkin, Ph. (ed.): The Oxford handbook of lexicography. Oxford, pp. 546–560.

Trier, J. (1931): Der deutsche Wortschatz im Sinnbezirk des Verstandes. Die Geschichte eines sprachlichen Feldes. Bd. I: Von den Anfängen bis zum Beginn des 13. Jahrhunderts. Heidelberg.

Trübner: Trübners Deutsches Wörterbuch. Im Auftrag der Arbeitsgemeinschaft für deutsche Wortforschung bearbeitet von Alfred Götze, vols. 1–5. Berlin 1939–1957.

Williams, R. (1976): Keywords. A vocabulary of culture and society. New York.

Contact information

Volker Harm
Zentrum für digitale Lexikographie der deutschen Sprache (ZDL), Akademie der Wissenschaften zu Göttingen
vharm@uni-goettingen.de

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Andrea Moshövel

SKATOLOGISCHER WORTSCHATZ IM FRÜHNEUHOCHDEUTSCHEN ALS KULTURGESCHICHTLICHE UND LEXIKOGRAPHISCHE HERAUSFORDERUNG

Abstract This paper deals with the lexicographic treatment of the evidently plenty and pervasive scatological vocabulary, that is vocabulary concerning the process and products of bodily excretion (especially feces), in the synchronical Early New High German Dictionary (FWB = Frühneuhochdeutsches Wörterbuch) from a dictionary user's view. Initially, different cultural concepts of scatology by Norbert Elias, Michail Bachtin and Mary Douglas among others and the term taboo are reflected. Subsequently, selected lexical items such as words with a primary scatological meaning (e. g. *drek*, *kot*, *scheisse*), concealing expressions (euphemisms, periphrases, metaphors, e. g. *sitzen*, *seine notdurft tun, bauernveiel*), and certain aspects within the polysemy of the verb *scheissen* are discussed, the latter on the one hand referring to a physical process with uncontrollable aspects and on the other hand denoting a deliberate action and functionalized as a fighting word during the reformation. Focussing on different positions of lexicographical information within the microstructure of the FWB, the surveillance shows that in a synchronical perspective Early New High German scatological vocabulary is a heterogeneous and complex phenomenon due to speaker, context and respectively semantic and pragmatic purposes.

Keywords Scatological vocabulary; Early New High German; Early New High German Dictionary (FWB); historical lexicography; historical lexicology; cultural history

1. Einleitung

In frühneuhochdeutschen Texten des 15./16. Jahrhunderts, insbesondere in komischen Texten wie Fastnachtspielen und Schwankerzählungen, aber auch in polemischen Äußerungen Martin Luthers stößt man nicht selten auf Skatologisches.[1] Die Enzyklopädie des Märchens ist eines der wenigen Handbücher, das eine ausführliche Definition zu „Skatologie" bietet:

> S[katologie] bezeichnet mündl[iche] und schriftl[iche] Äußerungen, die Ausscheidungsvorgänge oder produkte des (menschlichen) Körpers, bes[onders] Kot und –> Urin (–> Exkremente) oder Darmwinde (–> Furz) zum Gegenstand haben. Da die Wahrnehmung der Ausscheidungsprodukte bei den meisten Menschen spontanen Ekel auslöst, scheint das Exkrementelle in allen Gesellschaften (mehr oder weniger stark) tabuisiert, und sprachliche Äußerungen skatologischen Inhalts werden als obszön empfunden [...]. Dabei liegt die Schwelle der Empörung je nach Epoche, gesellschaftlicher Schichtzugehörigkeit etc. unterschiedlich hoch." (Gier 2007, c. 761)

Aus heutiger Sicht erscheint es selbstverständlich, dass skatologischer Wortschatz der Tabuisierung unterliegt und dass er (überwiegend negativ) emotional besetzt ist bzw. emotionale Reaktionen auslöst (Lazić/Mihaljević 2021, S. 647, n. 5; Mohr 2013, S. 5; 13 f. passim). Entsprechend macht die Integration skatologischen Wortschatzes in ein gesamtsprachbezogenes Wörterbuch, wie Radtke (1990, S. 1194) es für den sexuellen Wortschatz postuliert hat,

[1] Vgl. weiterführend Schmidt/Simon (2004, S. 109): „The German language has an unusually high elective affinity for scatological language."

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

diasystematische Markierungen erforderlich, d. h. Hinweise auf Wortverwendungsgrenzen und sprachkommunikative Grenzverletzungen (WLWF 2, S. 48–50, hier S. 49). Der von Gier (2007, c. 761) und vor allem in der englischsprachigen Tabuwörter-Forschung vielfach verwendete Begriff des „Obszönen" (Lazić/Mihaljević 2021, 643; Mohr 2003; 2013, S. 9 f.; 12; 17 f.) ist allerdings nicht unproblematisch, da er „moralische Komponenten" enthält und „kulturabhängig definiert" werden muss (Fährmann 2002, c. 179): „Eine Sache ist nie an sich obszön, sondern sie wird durch die Überschreitung subjektiv definierter Grenzen als solche aufgefaßt […]." (Ebd.)

In meinem Beitrag möchte ich Verwendungszusammenhängen des Skatologischen im Frühneuhochdeutschen und seiner lexikographischen Beschreibung in den bisher erschienenen Bänden einschließlich bislang noch unveröffentlichter Artikelentwürfe des Frühneuhochdeutschen Wörterbuchs (FWB) nachgehen. Damit wähle ich ein Wörterbuch, das in synchronischer Perspektive den Wortschatz des Frühneuhochdeutschen, d. h. der regionalen (oberdeutschen, mitteldeutschen, norddeutschen) und sozialen (schichten- und gruppenspezifischen) sprachlichen Varietäten des hochdeutschen Sprachraums von etwa der Mitte des 14. bis zur Mitte des 17. Jahrhunderts semantisch beschreibt.[2] Zunächst sollen einige theoretische und kulturgeschichtliche Voraussetzungen im Umgang mit dem Skatologischen geklärt und der Forschungsstand zum skatologischen Wortschatz im Frühneuhochdeutschen skizziert werden (Kap. 2). Anschließend sollen exemplarisch Formen und Funktionen des Skatologischen im FWB untersucht werden (Kap. 3), um abschließend mit Blick auf eine synchrone historische Perspektive, wie sie die Informationsaufbereitung im FWB bietet, Möglichkeiten und Aufgaben für die historische Lexikographie und ihren kulturgeschichtlichen Aussagewert zu formulieren (Kap. 4).

2. Forschungsgeschichtliche Annäherungen und Forschungsstand

In der soziologischen Forschung ist ein Wandel der Scham- und Peinlichkeitsschwelle vom Spätmittelalter bis heute und damit auch ein Wandel der Tabuisierung des Skatologischen postuliert worden. Nobert Elias (1977 [1939]) zufolge bildeten sich die feineren Sitten im Bereich des Fäkalverhaltens erst im Laufe des 16. Jahrhunderts und hierbei zunächst in der höfischen Gesellschaft und den städtischen Oberschichten aus, wie er insbesondere anhand von Benimmcodes zeigen kann (vgl. z. B. Elias 1977 [1939], Bd. 1, S. 177). An seinen wirkmächtigen, aus heutiger Sicht jedoch zu stark pauschalisierenden und aufgrund seiner teleologischen Ausrichtung nicht unproblematischen Ansatz knüpft beispielsweise Melissa Mohr (2003, 2013) in ihren Studien zum Obszönen im englischen Mittelalter und in der englischen Renaissance an, um differenzierend Lizenzen und Funktionen des Anstößigen in bestimmten Kontexten herauszuarbeiten, u. a. im Zusammenhang mit Beleidigungen und Komik oder Statuskonzepten und sozialen Hierarchien (vgl. Mohr 2003, S. 267–272; 2013, S. 106 f.; 162).

[2] Informationen zur Entstehung und Geschichte des Projekts bietet die Projekt-Homepage unter https://adw-goe.de/forschung/forschungsprojekte-akademienprogramm/fruehneuhochdeutsches-woerterbuch/ (Stand: 22.3.2022); zum Stand der gedruckten Ausgabe siehe https://adw-goe.de/forschung/forschungsprojekte-akademienprogramm/fruehneuhochdeutsches-woerterbuch/veroeffentlichungen/ (Stand: 22.3.2022); zu den Online verfügbaren Alphabetstrecken siehe https://fwb-online.de/content/verfuegbare-alphabetsstrecken (Stand: 22.3.2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Komik und ihre Wirkung unterliegen allerdings ebenfalls historischem Wandel. So existierte dem russischen Literaturtheoretiker Michail Bachtin (1969) zufolge in der Zeit des europäischen Spätmittelalters und der frühen Neuzeit neben der durch Kirche, Staat und Feudalherrschaft repräsentierten ernsten Kultur eine so genannte „mittelalterliche Lachkultur". Beschränkt auf die relative Freiheit einer „Insularität des Ausnahmezustands" (ebd., S. 34), die utopisch und auf den Karneval, d. h. auf die Festtage begrenzt war, wurde, so Bachtin, die Lachkultur von der „Wahrheit" der „grotesken Formen des Leibes" dominiert, d. h. von allen hervorstehenden und offen stehenden Körperteilen, in denen die „Grenzen zwischen Leib und Leib und Leib und Welt im Zuge eines Austausches und einer gegenseitigen Orientierung überwunden werden" (ebd., S. 17), und befreite nicht nur von der äußeren, sondern auch von der inneren Furcht, „von der in Jahrtausenden dem Menschen anerzogenen Furcht vor dem Geheiligten, dem autoritären Verbot, dem Vergangenen, vor der Macht" (ebd., S. 39).

Tatsächlich umfasst der Begriff des Tabus, unter dem man heute vor allem „Handlungen, Verhaltensweisen, bildliche Darstellungen, Themen und sprachliche Ausdrücke" versteht, „die in einer kulturellen Gemeinschaft von einer Mehrheit als nicht geduldet betrachtet werden" (Kocher 2009, c. 403), ursprünglich nicht nur „Verbotenes" und „Unreines", sondern auch „Geheiligtes":

> Der Begriff ‚T[abu]' stammt von polynesisch *tabu* oder *tapu* (bzw. *ta pu*) und wird im allgemeinen mit ‚geheiligt' oder ‚verboten' übersetzt, heißt aber lediglich ‚besonders markiert'. ‚T[abu]' ist, wer oder was über eine besondere Kraft verfügt und deshalb allein durch seine Existenz besondere Vorsicht oder Verhaltensregeln fordert. (Ebd.)

Tabubrüche, so Kocher (2009) weiter, sind stets „mit starken Emotionen verbunden" und müssen „nicht immer rational nachvollziehbar sein" (ebd., c. 405). Sie können u. a. „als ästhetische Effekte inszeniert werden, um Aufmerksamkeit zu erzeugen" und dabei fließende „Übergänge zu Schimpfwörtern und obszönen Ausdrücken" aufweisen. (Ebd.)

Mit Blick auf den besonderen Status von „Schmutz" konstatiert aus religionsethnologischer Sicht Mary Douglas, dass dieser „nur vom Auge des Betrachters aus" existiert (Douglas 1985 [1966], S. 12) und im Zusammenhang mit einem „System" und dessen Verwerfungen betrachtet werden muss:

> Schmutz ist […] niemals ein einmaliges, isoliertes Ereignis. Wo es Schmutz gibt, gibt es auch ein System. Schmutz ist das Nebenprodukt eines systematischen Ordnens und Klassifizierens von Sachen, und zwar deshalb, weil Ordnen das Verwerfen ungeeigneter Elemente einschließt. (Douglas 1985 [1966], S. 53)

„Unsauberes" wird von Douglas u. a. als das definiert, „was fehl am Platz ist", bzw. „das, was nicht dazugehören darf, wenn ein Muster Bestand haben soll" (ebd., S. 59). Als Symbole sowohl für Gefahren als auch für Kräfte (ebd., S. 159) sind Körperausscheidungen und Schmutz entsprechend widersprüchlich und komplex und können im Rahmen ritualisierten Handelns durchaus auch „zum Guten" genutzt werden:

> Die Gefahr, die bei einer Grenzüberschreitung droht, ist eine Entfesselung von Kräften. Die verletzlichen Randbereiche und aggressiven Kräfte, die die anerkannte Ordnung zu zerstören drohen, sind Ausdruck der Kräfte, die dem Kosmos innewohnen. Ein Ritual, das sie zum Guten einsetzen kann, ist im wahrsten Sinne kraftvoll. (Ebd., S. 210)

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Textsortenbezogene Bestandsaufnahmen speziell zum frühneuhochdeutschen skatologischen Wortschatz im Vergleich mit dem sexuellen Wortschatz, der sich jedoch in seiner Semantik und Pragmatik unterscheidet, bieten Müller (1988) für die Fastnachtspiele des 15. Jahrhunderts und Hanisch (1994) für die großen Schwanksammlungen des 16. Jahrhunderts. Müller (1988) zufolge ist „der Wortschatz der anal-skatologischen Obszönität" gegenüber „der überaus farbigen, bildhaften, facetten- und variantenreichen Metaphorik für in engerem Sinne Sexuelles […] wenig zahlreich und kunstvoll" (ebd., S. 196). Neben einer Vielzahl eigentlicher Ausdrücke führt er skatologische „Umschreibungen für den Kot" an, die ihm zufolge allerdings oft erst „im weiteren Zusammenhang deutlich" werden (ebd., S. 203) und vor allem „als Anlass" dienen, „den immer gleichen Scherz in aller Ausführlichkeit auszuschmücken" (ebd.). Hanisch (1994) verweist ebenfalls auf Schwierigkeiten des „Erkennen[s] von zweiten, anstößigen Bedeutungen" (ebd., S. 101) aus heutiger Sicht und postuliert, dass im Unterschied zum Bereich des Sexuellen, der mit dem „Witz des Verborgenen" spiele, der Bereich des Skatologischen durch „eigentliche Ausdrücke und deren Schockwirkung" gekennzeichnet sei (vgl. Hanisch 1994, S. 149). Seine Wortbeispiele zeigen dabei zugleich, dass im Frühneuhochdeutschen lexikalische Ausdrücke skatologischen Inhalts tropisch funktionalisiert werden können, um Nichtskatologisches auszudrücken bzw. auf Nichtskatologisches übertragen zu werden – und zwar offensichtlich in stärkerem Ausmaß als Ausdrücke aus dem Bereich des Sexuellen, zu denen er in der Funktion für Nichtsexuelles kaum Beispiele bietet (vgl. ebd., S. 155–158).

3. Skatologische Lexik im FWB – ein Streifzug

3.1 Schmutz und menschlicher Kot: zur Semantik von *drek*, [1]*kot*, [1]*mist*, *scheis* und *scheisse*

In den Fastnachtspielen des 15. Jahrhunderts und den Schwanksammlungen des 16. Jahrhunderts sind die häufigsten eigentlichen Ausdrücke für Exkremente *drek* und *kot* (vgl. Müller 1988, S. 202; Hanisch 1994, S. 106; 149). Ihre im FWB beschriebene Polysemie wird in der folgenden gekürzten Übersicht der Polysemie von [1]*mist* gegenübergestellt, da allen drei Wörtern gemeinsam ist, dass sie sich nicht nur auf menschlichen (und teils tierischen) Kot, sondern auch auf >Schmutz< allgemein sowie auf Nichtskatologisches beziehen, wenngleich mit unterschiedlichen Nuancierungen (siehe Tab. 1):

Ba.	***Drek***	**[1]*kot***	**[1]*mist***
1.	>Kot, Exkremente<; >Mist<; ütr.: >alles, was j. von sich gibt<; >Wertloses, Schlechtes, Abfall<	>Ausscheidung des Darms, Kot; Harn<	>Exkremente, Kot des Menschen und einiger Tiere< (Aspekte u. a.: Nutzen für alltagsbezügliche, darunter therapeutische Zwecke u. ä.)
2.	>Dreck, Schmutz; (Straßen) schlamm, Staub<; tropisch: >Nichts<	>Schmutz, Dreck, (Straßen) schlamm<	>Mist, aus dem Kot von Menschen und (vor allem:) von größeren Tieren; Gemisch zur Düngung des Bodens< (als Wirtschaftsgut); ütr. auch: >Förderung, Hilfe< sowie zeichenhaft für *schmachheit* o. Ä.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Ba.	*Drek*	[1]*kot*	[1]*mist*
3.	als Ort metaphorisierte Verhältnisse, in denen sich sowohl Schweine wie lasterhafte Menschen (gerne) aufhalten	>Erde, Lehm, Bauschutt<	>Kehricht, Abfall, Schmutz, Müll< (wie er sich auf Straßen, Plätzen etc. findet); mehrfach ütr.: >als Kehricht gedachte, weltverhaftete Sündenschicht im Menschen<
4.	Abweisungsformel und verächtlich machendes, teils dehumanisierendes, auch auf Tiere sowie konkrete und abstrakte Gegenstände bezogenes Schimpfwort		*mist* als Zeichen für >Gestank, Fäulnis, Dreck, Schmutz<; >Nichtigkeit<; teils mit der religiös motivierten Nuance: >Bodensatz, unterste Schicht natürlicher Verfallenheit, Verworfenheit, Wertlosigkeit<

Tab. 1: Zur Semasiologie von frnhd. *drek*, [1]*kot*, [1]*mist* (Ba. = Bedeutungsansätze)

Die schwierige und nicht immer eindeutig mögliche Abgrenzbarkeit von Schmutz allgemein und menschlichen Ausscheidungen in der jeweiligen Semantik von frnhd. *drek*, [1]*kot* und [1]*mist* legt verschiedene Erklärungsansätze nahe: Zum Ersten könnte sie darauf verweisen, dass in der Alltagswelt des Spätmittelalters und der Frühen Neuzeit der allgegenwärtige Schmutz immer auch mit Exkrementen jeglicher Art vermischt sein konnte bzw. auf diese Weise wahrgenommen wurde. Zum Zweiten könnte sie aber auch davon zeugen, dass es sich nicht um eigentliche Ausdrücke, sondern um (ehemalige) Euphemismen handelt (Euphemismen seien hier als „mildernde oder beschönigende Ersatzwörter" aufgefasst; Herchert 1996, S. 58). Im FWB-Belegmaterial lassen sich durchaus Hinweise finden, dass, wie Douglas (1985 [1986], S. 208) schlüssig ausgeführt hat, Exkremente, die im allgemeinen Schmutz aufgehen, als weniger bedrohlich wahrgenommen werden als der als solcher identifizierbare menschliche Kot:[3]

(1) Bischoff u. a., Steir. u. kärnt. Taid. 113, 13 (m/soobd., 1603): *es soll ein jeder purger* [...] *daß kott vor seinem hauß wekraumen.* [s. v. [1]*kot* 2]

(2) Sachs 20, 325, 5 (Nürnb. 1563): *Vom menschenkote her | Da wurd die archen also schwer, | Daß sie sanck und wolt untergehn.* [s. v. *menschenkot*]

In den in religiösen Kontexten auftretenden, vielfältigen tropischen und überwiegend abwertenden Bedeutungen von frnhd. *drek*, [1]*kot* und [1]*mist* für Nichtskatologisches im Sinne des Niedrigen, Wertlosen, Nichtigen, Marginalisierten, Ausgegrenzten, Verworfenen, Lasterhaften, Sündhaften, Schmählichen oder auch Bedeutungslosen scheint eine scharfe Abgrenzung zwischen der Bedrohlichkeit durch Schmutz allgemein und der Bedrohlichkeit durch menschlichen Kot speziell jedoch kaum möglich und daher nicht von Relevanz zu sein.[4]

[3] Alle im Folgenden verwendeten frühneuhochdeutschen Quellenzitate mit Orts- und Zeitangabe entstammen dem Korpus und dem Belegmaterial des FWB. Die bibliographischen Angaben zu den verwendeten Siglen finden sich im Quellenverzeichnis in FWB 1, 165–224; FWB 3, XV–XXI sowie unter https://fwb-online.de/content/bibliographie (Stand: 22.3.2022).

[4] Ausnahmen bilden seltene positive Übertragungen wie beispielsweise >Förderung, Hilfe< s. v. [1]*mist* 2.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Die Semantik der frnhd. Ausdrücke *scheis* >Darmwind< und *scheisse* >Durchfall< unterscheidet sich sowohl von den Verwendungsweisen von *drek*, [1]*kot* und [1]*mist* als auch vom modernen heutigen Gebrauch, was in der Mikrostruktur der entsprechenden Artikelentwürfe insbesondere anhand der angeführten Bedeutungsverwandten sichtbar wird (siehe Tab. 2):

Ba.	***scheis* (*der*)**	***scheisse* (*die*)**
1.	>(geräuschvoller) Darmwind<; mit Tendenz zum Phrasem auch derb bildhaft für >Nichts<. -- Bdv.: *fist, furz;* vgl. *bombart* 2, *gestank* 1, *gewind* (*der*), *komphart, stank.*	>Durchfall, Diarrhö; Ruhr<. -- Bdv.: auslauf 8, bauchflus, dünscheis, durchlauf 3, *strunt, unflat;* vgl. *durchgang* 3, *dysenterie*, rur 1, *stulgang.*
2.	phras.: *jm. einen scheis ausjagen / einjagen* >jm. fürchterliche Angst, einen großen Schrecken einjagen<.	

Tab. 2: Zur Semasiologie von frnhd. *scheis* und *scheisse* (Ba. = Bedeutungsansätze; Bdv. = Bedeutungsverwandte)

Die Belegzitatauswahl zeigt hierbei, dass sowohl frnhd. *scheis* als auch *scheisse* in den zeitgenössischen Wörterbüchern von Henisch und Dasypodius nicht diasystematisch markiert sind (Beispiele 3–4)[5] und dass sich beide Ausdrücke (und das gilt in gewisser Weise auch für das Phrasem *jm. einen scheis ausjagen / einjagen*) auf unwillkürliche bzw. unkontrollierte Körpervorgänge beziehen (Beispiele 5–7):

(3) Henisch 1315 (Augsb. 1616): *Furtzen / ein furtz oder scheiß lassen / pedere, crepitum reddere.* [s. v. *scheis* 1]

(4) Dasypodius 407v (Straßb. 1536): *der die Scheiß hat. Foriolus pe. cor.* [s. v. *scheisse*]

(5) Sachs 9, 269, 12 (Nürnb. 1555): *Als er sich thet so jheling bucken* | Da ließ er einen lauten scheiß. [s. v. *scheis* 1]

(6) Sachs 23, 93, 25 (1556): *Ich wil dem altn ein-jagn ein schais | Und in setzen in ein angst-schwais.* [s. v. *scheis* 2]

(7) Chron. Strassb. 8, 346, 13 (els., A. 15. Jh.): [*Vespasianus*] *starp an der schysse also er was 69 jor alt.* [s. v. *scheisse*]

3.2 Verhüllendes Sprechen über Exkrementelles: Euphemismen, Periphrasen, metaphorische Ausdrücke

Neben den in 3.1 behandelten eigentlichen Ausdrücken für Skatologisches bietet das Frühneuhochdeutsche auch verhüllende Euphemismen und Umschreibungen bzw. Periphrasen (als „umschreibende Ausdrucksweise verstanden, die mit mehreren Wörtern auseinandersetzt, was mit einem oder mit weniger gesagt werden kann"; Ueding/Steinbrink 1986, S. 268;

[5] Vgl. zum heutigen Gebrauch beispielsweise unter Duden online (2022) die Bedeutungsangaben und diasystematischen Markierungen „derb" s. v. *Scheiße, scheißen* bzw. „salopp abwertend" s. v. *Scheiß*: https://www.duden.de/rechtschreibung/Scheisze (Stand: 18.5.2022); https://www.duden.de/rechtschreibung/scheiszen (Stand: 18.5.2022); https://www.duden.de/rechtschreibung/Scheisz (Stand: 18.5.2022); vgl. auch Nübling/Vogel (2004).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Herchert 1996, S. 58). Ein gutes Beispiel findet sich in Luthers Bibelübersetzung mit dem Euphemismus *sitzen* und den Umschreibungen *zur not hinausgehen* sowie *was von jm. gegangen ist*:[6]

(8) Luther. Hl. Schrifft. 5. Mose 23, 13 f. (Wittenb. 1545): *VND du solt aussen fur dem Lager einen Ort haben / da hin du zur not hinaus gehest. Vnd solt ein Scheufflin haben / vnd wenn du dich draussen setzen wilt / soltu da mit graben / vnd wenn du gesessen bist / soltu zuscharren was von dir gangen ist.*

Für Elias' These von der zunehmend tabuisierenden Normierung des Skatologischen im 16. Jahrhundert könnte hierbei sprechen, dass im Unterschied zu Luther und anderen – wie Luther verhüllend umschreibenden – zeitgenössischen Bibel-Übersetzungen des 16. Jahrhunderts in der ersten vollständigen deutschen Vulgata-Übersetzung, der 1466 in Straßburg gedruckten Mentelin-Bibel sowie ihrer Revision, Augsburg 1475, die kaum verhüllenden lexikalischen Ausdrücke *gestank* (Kurrelmeyer, Dt. Bibel 4, 206, 2 ff. [Straßb. 1466]: „*bedeckst den gestanck mit erden den du hast deroffent*") und *kot* (ebd. Var. 1475: „*bedeck dein kot mit der erde mit dem du dich hast enthoebt*") verwendet werden.[7]

Als Euphemismen, die in Fachtexten der Medizin bzw. fach- und wirtschaftsbezüglichen Texten für Exkrementelles gebräuchlich sind, lassen sich im FWB beispielsweise *materie* 4 für >Exkremente<, *stul* 3, eigentlich >Toilettenstuhl<, mit anschließbaren Metonymien wie >Stuhlgang< (als relevant für die Gesundheit generell); >Exkremente, Kot<; >Durchfall< u. a. sowie das Kompositum *stulgang* finden, das allerdings (ähnlich wie *scheisse*) mit dem Hinweis auf die in der Mehrzahl der Belege auftretende Bedeutung >Durchfall, krankhafte Ausscheidung (als Vorgang) des Darms; Ruhr< semantisch anders nuanciert erscheint als heute.

Vorgang und Handlung der Defäkation werden zudem vielfach mit dem – höchst polysemen bzw. unterspezifizierten – Verb *tun* in Bedeutungsansatz *tun* 17 >seine Notdurft verrichten; Exkremente ausscheiden< (Beispiele 9–11) und dem Ausdruck *notdurft* 2 in der Bedeutung >Notdurft, Entleerung des Darms< ausgedrückt, der in der phrasematischen Wendung *seine notdurft tun* auch wiederholt im Kotext zusammen mit dem Verb *scheissen* steht und (vielleicht deshalb) im FWB mit der diasystematischen Angabe „sowohl verhüllend wie neutral und derb gebraucht" versehen ist (Beispiele 12; 13):

(9) Banz, Christus u. d. minn. Seele 88 (alem., 1. H. 15. Jh.): *Villicht het es [das kind] under sich geton: | So müstist denn wúschen [...] gon.* [s. v. *tun* 17]

(10) Luther. 38, 559, 24 (1538): *[Der Teuffel] isset gerne niedliche bissen und thut gern an reine orter, denn er helt seinen unflat fur thesem und balsam.* [s. v. *tun* 17]

(11) Mieder, Lehmann. Flor. 862, 5 (Lübeck 1639): *Clauß sahe das Fůrstliche Frawenzimmer / als sie im Garten spatzieren giengen / jhre Notturfft thåten / hernach wolt ers auch so machen / hůllet seinen Rock vnnd Hosen vmb sich / bruntzt vnd thet alles voll / als ob [...].* [s. v. *notdurft* 2]

(12) Maaler 308r (Zürich 1561): *Sein Notturfft thůn od' scheyssen. [...]. Sein Notturfft in ein guldin becke thůn.* [s. v. *notdurft* 2]

[6] Die Angaben von Syntagmen erfolgen nach den Konventionen des FWB; vgl. zum kulturgeschichtlichen Aussagewert von Syntagmenangaben allgemein Lobenstein-Reichmann (2002).

[7] Ich danke meiner Kollegin Dr. Carola Redzich für ihre diesbezügliche Recherche im FWB-internen Bibelarchiv.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

(13) Chron. Strassb. 369, 2 (els., A. 15. Jh.): *wann der keyser Constancius fůr eines moles durchehtende die cristenheit, und also er wolte sine notdurft tůn, do scheis er sin ingeweide mit dem bothe* [siehe [1]*bacht* 1 > Schmutz, Dreck<] *herus, das er ze stunt starp.* [s. v. *notdurft* 2]

Kontrovers wird in den Quellen insbesondere die Defäkation des Jesuskindes diskutiert. Der Erzählung vom Leermachen der Windeln steht an anderer Stelle die Aussage entgegen, dass am Jesuskind nie *unflat* oder *unreine ding* wahrgenommen worden seien, sondern man es immer *trucken* (hier entsprechend im Sinne von >sauber; frei von flüssigen Ausscheidungen<; s. v. *trucken* 10) vorgefunden hätte:

(14) Adrian, Saelden Hort 1569 (alem., Hss. E. 14./15. Jh.): *vil togenlichen blik* | zů *siner notdurft, nim si war* | *und maches windellinen bar* (Kontext: Pflege des Jesuskindes). [s. v. *notdurft* 2]

(15) Päpke, Marienl. Wernher 595 (halem., v. 1382): *An im* [Jesus, dem Kleinkind] *und aller siner wåt* | *Wart nie gesechen unflåt,* | *Måsen, fleken kaine* | *Noch ander ding unraine.* | *Truken, schôn man es vand* | *Als man es laite von der hand.* [s. v. *trucken* 10]

Verschiedene Ausprägungen zeigen in frühneuhochdeutschen Texten konventionelle und kontextgebundene Metaphern (zu Klassifizierung und jeweiligen Merkmalen vgl. allgemein Herchert 1996, S. 58 ff.). Auf jeweils einem Kontrast beruhen beispielsweise *rose* für >Stuhlgang< und *röseln* (V.) für >‚Röschen' (d. h. Stuhlgang) ausscheiden< (s. v. *rose* 1; Beispiel 16), der Ausdruck [1]*griebe*, mit dem in der eigentlichen Bedeutung eines Nahrungsmittels (>Speckwürfel<) in Verbindung mit *scheishaufen* ein (aus heutiger Sicht) besonders drastischer Effekt derber Komik entsteht (Beispiel 17), oder [1]*könig* 4, dessen Bedeutungsangabe >Kot-; Dreckhaufen< mit dem Hinweis „bei Umsetzung des Würde-Gedankens in sein Gegenteil anschließbar an 2" eine (aus heutiger Sicht) nachvollziehbare semantische Motiviertheit des metaphorischen Verwendungszusammenhangs nahelegt (Beispiel 18):

(16) Hampe, Ged. v. Hausrat 4, 25, 10 (Straßb. um 1514): *Das Stũlin dar vff es [kindel] dan Rôßlen sol* | *Die selben roßen schmecken* [>riechen<] *nit fast wol.* [s. v. *rose* 1]

(17) Lichtenstein, Lindener. Katzip. 302 (o. O. 1558): *[des mägdleins můtter] erwuscht ein grossen scheyßhafen [...] und schlegt den gůten kürschner* für sein *schnautzen, das im die griffen* [>Grieben<] *an der goschen kleben, und die wurst recht briete.* [s. v. [1]*griebe*]

(18) Fischer, Folz. Reimp. 18, 131 (Nürnb. um 1520): Sol ich ein hafen [>Topf<] nötigs han, | Find ich vol prunczwassers [>Urin<] stan | Und einen künig unten drin. [s. v. [1]*könig* 4]

Den Übergang von kontextueller zu konventioneller Metapher illustriert besonders eindrücklich der Ausdruck *bauernveiel*. Er geht auf den in der Neidhart-Tradition „in Wort und Bild" (Wachinger 2010 [2006], S. 679) weit verbreiteten Veilchenschwank zurück: Der Minnesänger Neidhart entdeckt das erste Frühlingsveilchen (mhd. *vîol*, frnhd. *veiel*) und bedeckt es mit seinem Hut, um die Herzogin herbeizuholen und es ihr zu zeigen. Als diese mit ihrem Gefolge herbeikommt und der Hut gelüftet wird, muss er jedoch feststellen, dass ihm die Bauern einen Streich gespielt und während seiner Abwesenheit einen Kothaufen unter dem Hut platziert haben. In der Überlieferung wird dieser beispielsweise mit dem verhüllenden Latinismus *merdum* bezeichnet, u. a. in einem Fastnachtspiel von Hans Sachs:[8]

(19) Sachs 17, 198, 16 (Nürnb. 1557): *Als die fürstin den merdrum fand,* | *Bestund Neydhart mit spot und schand.* Ebd. 201, 9: *Thet darnach den merdrum auffdecken,* | *Der feyhel würd ir nit wol schmecken.* [s. v. *merdum*]

[8] Darüber hinaus bietet die Überlieferung z. B. (nicht deutbares) *sor* bzw. *sorge*, *kunter* >Ungetüm<; >Kot, Dreck< (vgl. FWB s. v. *kunter* 2; 3) oder auch den beredten Verzicht auf eine Benennung (vgl. Wachinger 2010 [2006]), S. 86–95; S. 679–685).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

In Johannes Paulis Schwanksammlung *Schimpf und Ernst* (1522) hingegen wird der Ausdruck *bauernveiel* losgelöst von dem konkreten Kontext des Veilchenschwanks verwendet. Der Verfasser macht sich allerdings analog zur Handlungsebene des Veilchenschwanks mit dem Euphemismus *hofieren* im Sinne von >Exkremente ausscheiden< den Kontrast von *hof* und *bauer* für einen komischen Effekt auf der Ebene eines intertextuell anspielungsreichen Wortspiels zunutze:

(20) Bolte, Pauli. Schimpf u. Ernst 1, 363, 28 (Straßb. 1522): *Der Pfarrer sprach: ‚Es sol gelten', und hofiert in die Kirchen, und satzt ein großen Baurenvigel.* Ebd. 364, 12: *Der Abenthůrer stůnd uff und hofiert an des Pfaffen Bett ein grossen Baurenfigel und ein große Lachen* [>Urinlache<]. [s. v. *bauernveiel*]

Die Beispiele sollen ausreichen, um einen Eindruck von der Spannbreite und den unterschiedlichen ästhetischen Funktionen des verhüllenden Sprechens über Exkrementelles in verschiedenen Textsorten des Frühneuhochdeutschen zu vermitteln: Während Euphemismen, Periphrasen, Metaphern und ihre Verwendung in Bibelübersetzungen, Fach- und Wirtschaftstexten sowie erbaulichen bzw. religiös-didaktischen Texten einen tabuisierenden Sprachgebrauch implizieren, werden sie hingegen in den auf Lachwirkung zielenden komischen literarischen Gattungen zur Ausschmückung, zur Erzielung besonders drastischer Effekte skatologischer Komik oder als Ausweis anspielungsreichen Wortwitzes genutzt.

3.3 Physiologischer Vorgang und machtvolles Handeln: zur Funktionalisierung des Verbs *scheissen*

Das semasiologische Feld des u. a. in den Fastnachtspielen „[s]ehr viel häufiger als die entsprechenden Substantive" (Müller 1988, S. 200) belegten Verbs *scheissen* ist im Artikelentwurf des FWB in drei Bedeutungsansätze gegliedert (siehe Tab. 3):

scheissen 1	>koten, den Darm entleeren< (als physiologischer Vorgang); auch speziell auf Diarrhö bezogen; häufig subst. verwendet; offen zu 2; 3.
scheissen 2	an einem / einen, nicht dafür bestimmten Ort koten; etw. (mit Kot) verunreinigen, besudeln< (als anstößige, verwerfliche Handlung); mit Tendenz zur Phrasematisierung Geringschätzung, Verachtung und Aggressivität gegenüber e. P. / e. S. ausdrückend; meist derb, auch zur Bildung eines affektgeladenen, negationsverstärkenden Kompositionselements in Schimpfwörtern und Flüchen verwendet.
scheissen 3	>Darmwinde, Blähungen entweichen lassen<; im Orientierungsfeld mit anderen geräuschvollen 'Unarten' wie *garzen, husten, jucken* 3, *koppen, kratzen* 1, *kreisten, ratzen, rotzen.*

Tab. 3: Zur Semasiologie von frnhd. *scheissen*

Auffällig ist vor allem die semantische Unterscheidung zwischen den nur begrenzt steuerbaren, physiologischen Vorgängen der Defäkation (*scheissen* 1) und Blähung (*scheissen* 3) und einem willkürlich intendierten, aktiven und aggressiv-zerstörerischen Handeln (*scheissen* 2).[9] Innerhalb der Mikrostruktur des Artikels illustrieren die unter Bedeutungsansatz 2

[9] Diese Unterscheidung ist im FWB übrigens auch für den reflexiven und transitiven Gebrauch von frnhd. *bescheissen* 2 zu beobachten, wie die Syntagmen *sich (vor angst / lachen / leide) bescheissen* gegenüber *jn. mit drek bescheissen* verdeutlichen können. Auf weitere Besonderheiten der Semantik

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

verzeichneten Phraseme, Syntagmen, Wortbildungen und Belegzitate, dass es sich auch um eine Semantisierung und Funktionalisierung als Kampfbegriff im Zuge der Reformation handelt (siehe Tab. 4; vgl. mit weiterem Belegmaterial auch Schmidt/Simon 2004):

	scheissen 2
Phras.	*in etw.* (z. B. *in den tauf*) *scheissen* >auf nichts Wert legen, etw. verächtlich behandeln< (als abfälliger Kommentar der Geringschätzung teils unter Strafe gestellt; vgl. Rwb 12, 405); *in das eigene nest scheissen* >sich selbst Schaden zufügen<; *(jm.) in die hände scheissen* >jn. handlungsunfähig machen<; *jm. auf das / in das maul scheissen* >jn. bestrafen<; *jm. ins mus scheissen* >jm. etw. verderben, js. Pläne durchkreuzen<; ⌜*da scheis der esel dahin!; da scheis ein hund ein!* u. Ä.⌝ jeweils als Ausdruck der Verachtung; *scheis in die bruch und hänge sie an den hals* >Mach', was du willst!<.
Synt.	*j. einen drek* [wohin], *jm. einen drek an den weg, auf die nase, in die augen s., der bapst die artikel der römischen kirche s.*; *der papst seines dreks die ganze welt vol s.*; *j. jm. in seinen zorn s.*; *j.* [wohin] (z. B. *in den senf, in die kirche, das bet*) *s., grob s.*
Wbg.	*scheis|ban, scheisdrek,* ⌜*scheispfaffe, scheispoet* (dazu bdv.: *arshummel*), *scheisser*⌝ jeweils als Schimpfwort verwendet.
Bz.	Luther. WA 47, 425, 11 (1537/40): *der [Bapst] ist des Teufels bischoff und der Teufel selbst, ja der Dreck, den der Teufel in die kirche geschissen hat.* Ebd. 466, 12: *man hat mussen glauben an die Artickel der Romisschen kirchen, die der Bapst geschissen hat.* Ebd. 49, 276, 37 (1543): *ich scheus euch ihn euren zorn.* Schade, Sat. u. Pasqu. 2, 67, 3 (o. O. 1524): *ich schiß eim in das wölfisch maul, der eim ein berg verhieß und künt im nit ein stein davon reichen.*

Tab. 4: Zur Mikrostruktur von *scheissen* 2 (Phras. = Phraseme.; Synt. = Syntagmen; Wbg. = Wortbildungen; Bz = Belegzitate)

Mit der Funktionalisierung zum Kampfbegriff kann etwas, das als geheiligt, ehrbar u. Ä. erachtet wird, verunreinigt und wertlos gemacht werden: Dabei kann der Ausdruck zum einen beleidigend und diffamierend als handlungszuschreibende Markierung eingesetzt werden, mit der die aus Sprechersicht zerstörerischen Machenschaften der Repräsentanten eines falschen Systems (*babst, teufel*) entlarvt werden. Zum anderen kann aber auch auf ihn zurückgegriffen werden, um sich selbst auf der Basis eines körperlichen Garanten von Wahrheit mit dem aggressiv aufgeladenen skatologischen Fluch- und Scheltwort als Waffe[10] kraft- und machtvoll handelnd gegen jemanden oder etwas, z. B. gegen Widersacher und das (aus Sprechersicht) falsche System, das sie repräsentieren, in Stellung zu bringen.[11]

von *bescheissen*, bei der sich analog zu frnhd. *drek*, [1]*kot*, [1]*mist* Handlungen allgemeinen Beschmutzens (s. v. *bescheissen* 1) und des speziellen Beschmutzens mit menschlichem Kot (s. v. *bescheissen* 2) nur schwer voneinander abgrenzen lassen und der übertragene Gebrauch auf Nichtskatologisches im Sinne von >jn. betrügen, überlisten, hinters Licht führen< (s. v. *bescheissen* 3) seit dem 15./16. Jahrhundert reich belegt ist, kann hier aus Platzgründen nicht eingegangen werden.

[10] Röcke (1987, S. 210) spricht im Zusammenhang mit dem skatologischen Eulenspiegel von der „apotropäischen Kraft der obszönen Gebärde", was sich auch auf Sprachgebärden beziehen lässt.

[11] Zu den Anschlussmöglichkeiten der theoretischen Ansätze bzw. Analysen von Bachtin (1969), Douglas (1985 [1966]) oder auch Mohr (2003) an diesen Befund vgl. Kapitel 2.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

4. Resümee

Zusammenfassend lässt sich feststellen, dass sich Formen und Funktionen skatologischer Lexik im Frühneuhochdeutschen von ihrem heutigen Gebrauch deutlich unterscheiden, dass hierbei jedoch die aus diachroner Perspektive immer wieder gestellte Frage nach einer genaueren Bestimmung des Tabuisierungsgrades als geringer oder ähnlich im Vergleich zu heute und/oder anderen Epochen zu kurz greift. Ein synchron angelegtes semantisches Wörterbuch wie das FWB vermag Benutzerinnen und Benutzern auch zu zeigen, wie heterogen und differenziert Gegenstände innerhalb einer Epoche sprachlich verhandelt und verfasst werden, wodurch die auf diachroner Ebene gewonnenen Ergebnisse ergänzt und zu einem gewissen Grad relativiert werden können. Nur eine entsprechende, von philologischer Tiefenschärfe und hermeneutischem Problembewusstsein geprägte (und damit letztlich nicht von Maschinen leistbare), auf die Handlungssemantik fokussierte Beschreibungsleistung, wie sie sich in der Mikrostruktur der FWB-Artikel spiegelt, kann zutage fördern, wie uneinheitlich und differenziert sich innerhalb der Epoche des Frühneuhochdeutschen der kommunikative Umgang mit dem (heute immer noch anrüchigen?) Skatologischen gestaltet. In Abhängigkeit von zeit-, raum- und textsortenabhängigen Faktoren spiegelt er unterschiedliche kommunikative Interessen wider, die sich zwischen Tabuisierung, Semantisierung und Funktionalisierung bewegen und manchmal sogar erst im (außergewöhnlichen) Einzelbeleg sichtbar werden (vgl. oben u. a. Beispiel 15). Damit leistet ein synchron angelegtes Sprachstadienwörterbuch wie das FWB Grundlagenforschung, die für das, was uns auch heute in der Gegenwart an kulturellen und sprachlichen Phänomenen bewegt (z. B. die Frage, was in welchen Kontexten als anstößig und verletzend gilt und wie es funktioniert), von Relevanz ist und an die entsprechend Wissenschaftlerinnen und Wissenschaftler verschiedener Disziplinen mit ihren jeweiligen Frage- und Erkenntnisinteressen anknüpfen können.

Literatur

Bachtin, M. (1969): Literatur und Karneval. Zur Romantheorie und Lachkultur. Aus dem Russischen von A. Kaempfe. München.

Douglas, M. (1985 [1966]): Reinheit und Gefährdung. Eine Studie zu Vorstellungen von Verunreinigung und Tabu. Aus dem Amerikanischen übersetzt von B. Luchesi (Originalausgabe: Purity and danger. London 1966). Berlin.

Duden online (2022): https://www.duden.de/woerterbuch (Stand: 18.5.2022).

Elias, N. (41977 [1939]): Über den Prozess der Zivilisation. 2 Bde. Frankfurt a. M.

Fährmann, S. (2002): [Art.] Obszönitäten. In: Brednich, R. W. et al. (Hg.): Enzyklopädie des Märchens. Handwörterbuch zur historischen und vergleichenden Erzählforschung. Bd. 10, S. 178–183.

FWB (1986 ff.): Goebel, U./Lobenstein-Reichmann, A./Reichmann, O. (Hg.) (1986 ff.): Frühneuhochdeutsches Wörterbuch. Seit 2013 im Auftrag der Akademie der Wissenschaften zu Göttingen. Berlin et al. https://fwb-online.de/ (Stand: 23.3.2022).

Gier, A. (2007): [Art.] Skatologie. In: Brednich, R. W. et al. (Hg.): Enzyklopädie des Märchens. Handwörterbuch zur historischen und vergleichenden Erzählforschung. Bd. 12. Berlin/New York, S. 762–766.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Hanisch, M. (1994): Das Obszöne in den Schwanksammlungen des 16. Jahrhunderts. Motive und Wortschatz bei Wickram und den nachfolgenden Autoren Frey, Montanus, Lindener und Schumann. Diplomarbeit zur Erlangung des akademischen Grades Magister phil. Wien.

Herchert, G. (1996): „Acker mir mein bestes Feld". Untersuchungen zu erotischen Liederbuchliedern des späten Mittelalters. Mit Wörterbuch und Textsammlung. Münster/New York.

Kocher, U. (2009): [Art.] Tabu. In: Ueding, G. (Hg.): Historisches Wörterbuch der Rhetorik. Bd. 9. Tübingen, S. 403–409.

Lazić, D./Mihaljević, A. (2021): Stereotypes and taboo words in dictionaries from a diachronic and a synchronic perspektive – the case study of Croatian and Croatian Church Slavonic. In: Gavriilidou, Z./Mitits, L./Kiosses, S. (Hg.): Proceedings of XIX EURALEX Congress: Lexicography for Inclusion. Vol. II. Thrace, S. 643–653.

Lobenstein-Reichmann, A. (2002): Die Syntagmenangabe – ein Stiefkind der Bedeutungslexikographie. In: Ágel, V. et al. (Hg.): Das Wort. Seine strukturelle und kulturelle Dimension. Tübingen, S. 71–87.

Mohr, M. (2003): Defining dirt: three early modern views of obscenity. In: Textual Practice 17 (2), S. 253–275.

Mohr, M. (2013): Holy sh*t. A brief history of swearing. Oxford.

Müller, J. (1988): Schwert und Scheide. Der sexuelle und skatologische Wortschatz im Nürnberger Fastnachtspiel des 15. Jahrhunderts. (= Deutsche Literatur von des Anfängen bis 1700). Bern u. a.

Nübling, D./Vogel, Marianne (2004): Fluchen und Schimpfen kontrastiv. Zur sexuellen, krankheitsbasierten, skatologischen und religiösen Fluch- und Schimpfwortprototypik im Niederländischen, Deutschen und Schwedischen. In: Germanistische Mitteilungen 59, S. 19–33.

Persels, J./Ganim, R. (2004): Introduction. Scatology, the last taboo. In: Persels, J./Ganim, R. (Hg.): Fecal matters in early modern literature and art. Studies in scatology. (= Studies in European Cultural Transition 21). Aldershot, S. xiii–xxi.

Radtke, E. (1990): Das Wörterbuch des sexuellen Wortschatzes. In: Hausmann F. J. et al. (Hg.): Wörterbücher. Dictionaries. Dictionnaires. Ein internationales Handbuch zur Lexikographie. 2. Teilbd. (= HSK 5/2). Berlin/New York, S. 1190–1199.

Röcke, W. (1987): Kollektive Mentalität und Individualisierung. Probleme einer historischen Poetik des Ulenspiegel. In: Schöttker, D./Wunderlich, W. (Hg.): Hermen Bote. Braunschweiger Autor zwischen Mittelalter und Neuzeit. Wiesbaden, S. 207–218.

Schmidt, J./Simon, M. (2004): Holy and unholy shit: the pragmatic context of scatological curses in early German reformation satire. In: Persels, J./Ganim, R. (Hg.): Fecal matters in early modern literature and art. Studies in scatology. (= Studies in European Cultural Transition 21). Aldershot, S. 109–117.

Ueding, G./Steinbrink, B. (21986): Grundriß der Rhetorik. Geschichte, Technik, Methode. Stuttgart.

Wachinger, Burghart (Hg.) (2010 [2006]): Lyrik des späten Mittelalters. Frankfurt a. M./Berlin.

WLWF (2010–20): Wiegand, H. E./Beißwenger, M./Gouws, R. (Hg.) (2010–2020): Wörterbuch zur Lexikographie und Wörterbuchforschung. Mit englischen Übersetzungen der Umtexte und Definitionen sowie Äquivalenten in neun Sprachen. Berlin/Boston.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Kontaktinformationen

Andrea Moshövel
Akademie der Wissenschaften zu Göttingen, Arbeitsstelle Frühneuhochdeutsches Wörterbuch (FWB)
amoshoe@gwdg.de

Danksagung

Der vorliegende Beitrag ist im Rahmen des Projekts Frühneuhochdeutsches Wörterbuch (FWB), finanziert von der Akademie der Wissenschaften zu Göttingen, entstanden.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Historical Lexicography: Romance and Other Languages

EURALEX XX 22
XX EURALEX INTERNATIONAL CONGRESS
DICTIONARIES AND SOCIETY
12-16 July 2022, Mannheim, Germany

Maria Arapopoulou/Georgios Kalafikis/Dimitra Karamitsou/
Efstratios Sarischoulis/Sotiris Tselikas

"VOCABULA GRAMMATICA": THREADING A DIGITAL ARIADNE'S STRING IN THE LABYRINTH OF ANCIENT GREEK SCHOLARSHIP

Abstract An ongoing academic and research program, the "Vocabula Grammatica" lexicon, implemented by the Centre for the Greek Language (Thessaloniki, Greece), aims at lemmatizing all the philological, grammatical, rhetorical, and metrical terms in the written texts of scholars (philologists and scholiasts) who curated the ancient Greek literature from the beginning of the Hellenistic period (4th/3rd c. BC) until the end of the Byzantine era (15th c. AD). In particular, it aspires to fill serious gaps (a) in the study of ancient Greek scholarship and (b) in the lexicography of the ancient Greek language and literature. By providing specific examples, we will highlight the typical and methodological features of the forthcoming dictionary.

Keywords Humanities; digital lexicography; specialized dictionary; Ancient Greek language; Ancient Greek scholarship

1. Introduction

Ancient Greek scholarship – γραμματική τέχνη in Greek, *ars grammatica* in Latin – refers to two organically linked spheres: on the one hand, to all literary works, their understanding, reconstruction, and interpretation, while on the other to the origins, structure, and functions of language as an autonomous tool for their creation. In that sense, *scholarship* relates, as an independent science, both to commentaries on literary texts (*hypomnemata*, *scholia*) and to exegetic grammatical treatises (Montanari 2011, pp. 11–13; Novokhatko 2015, p. 4; Matthaios 2015, p. 197).

Concerning the historical setting and development of the Greek scholarship, its earliest roots trace back to the 5th and 4th centuries BC during the Classical Age (479-323 BC); at that time, philosophers like Plato and Aristotle began to analyze the Greek language systematically. However, it was in the 3rd century BC Ptolemaic Alexandria, during the Hellenistic period (323-30 BC), when pioneering Greek scholars of its then world-famous academic institution, the Alexandrian Museum and Library, finally laid the solid foundations and disseminated Greek scholarship as a distinct scientific field. The field was flourishing from the 1st century BC through the start of the 6th century AD (Roman Imperial period and Late Antiquity). In the meantime, the scholarly tradition of the Greek East diffused into the Latin West; a novel School of Latin grammarians (*Grammatici Latini*) appeared and gradually developed in parallel to the standing Greek counterpart. From the 6th to the 15th century AD, the medieval Byzantine scholars significantly contributed to the survival and transmission of ancient Greek scholarship across Europe and beyond.

Despite having its starting point in Classics, interdisciplinarity characterizes the study of ancient Greek scholarship, thus expanding its perspective in other fields, such as modern literary theory, linguistics, rhetoric, and even philosophy. During the last decades, the an-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

cient Greek scholarship has been at the center of intensive and multidimensional research activity, decisively encouraged by modern editions of ancient or medieval commentaries and other related works.

2. A basic description of the "Vocabula Grammatica" dictionary

Having been designed from the beginning as a digital database using Drupal 6.38 as its Content Management System (CMS), its macro- and microstructure make salient its characteristics: *Vocabula Grammatica* is a specialized, linguistic, historical, multilingual, and polyphonic dictionary. It is implemented by the Centre for the Greek Language (Thessaloniki, Greece), under the supervision of Prof. A. Rengakos in collaboration with Prof. Franco Montanari (University of Genoa).

2.1 Macrostructure: term selection and ordering

Polyphony is a primary characteristic of the wordlist, which was composed by sorting out and cross-checking various indexes of critical editions or special studies (such as Grammatici Graeci II–IV; Erbse VI–VII (1983, 1988); Martin 1974; Meijering 1987; Keizer 1995; Lausberg 1998; Nünlist 2009), lexica and glossaries (such as Ernesti 1795; Bécares Botas 1985; Dickey 2007; Anderson 2000; Urrea Méndez 2003; Fenoglio 2012), falling under the determining fields of ancient Greek scholarship. This stage, although time-consuming, was a condition sine qua non for assuring the comprehensiveness of the dictionary. Hence, a wordlist of some 7.000 terms has been alphabetically elaborated. Nearly 2.000 grammatical terms, their derivatives, compounds, and other related terms are being currently lemmatized: e.g., συλλαβή-συλλαβικός-πολυσύλλαβος, πτώση-πτωτικός-ἄπτωτος-μονόπτωτος. Frequency was not the decisive criterium of the lemmatization because the specialized *Vocabula Grammatica* "aims at considerably higher terminological coverage" (Bergenholtz/Tarp (eds.) 1995, p. 90) than general dictionaries.

2.2 Dictionary material: the corpus and the canon

The dictionary aims at recording all the relevant terminology attested in scholarly works: grammatical and rhetorical treatises, lexica, commentaries and scholia, and works of textual criticism. Chronologically, our corpus spans over 20 centuries, from the 5th century BC to the 15th century AD. It includes:

a) early attestations from the Classical Age;
b) the scholars of the Early Hellenistic/Alexandrian period (3th-2nd c. BC), mainly librarians at Alexandria and Pergamon, such as Aristophanes of Byzantium and Aristarchus of Samothrace;
c) the scholars of the Late Hellenistic and the Roman imperial periods (2nd c. BC-5th c. AD), among whom the grammarians Dionysius Thrax, Apollonius Dyscolus and his son Aelius Herodianus prevail, as well as notable rhetoricians, such as Dionysius of Halicarnassus and Hermogenes of Tarsus; occasionally, Latin grammarians are also quoted, when a Greek term appears in Latin transcription.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

d) the scholars of the Byzantine period (5th-15th c. AD), among whom Hesychius, Georgius Choeroboscus, Photius, Eustathius of Thessalonica plus various etymological lexica merit a special mention.

(For all the above-mentioned cf. Sluiter 1990; Robins 1993; Dickey 2007; De Jonge 2008; Matthaios/Montanari/Rengakos (eds.) 2011; Matthaios 2014; Montanari/Matthaios/Rengakos (eds.) 2015; Tikkanen Westin 2018; Montanari (ed.) 2020).

2.3 Microstructure

The microstructure of our dictionary divides into three distinct but interrelated parts.

2.3.1 The Introductory Lemmatical Structure

In principle, *Vocabula Grammatica* is a linguistic dictionary. The headword appears according to received lexicographical practices for Ancient Greek (under **Lemma** (Fig.1)). Namely, for verbs, 1st person ind. pres. act. (ἀδρύνω; ἀθετέω -ῶ), or pass. if the term attests only to this voice (ἀναζωγραφέομαι -οῦμαι); for nouns, nom. and gen. sg. with the article (ἀναγωγή -ῆς, ἡ); for adjectives, participles, and verbal adjectives, nom. sg. in all genders (ἀβαρβάριστος, ἀβαρβάριστον; ἀδρός, ἀδρά, ἀδρόν). Variant spellings or forms are recorded under the same headword, where the most frequent appears first (δισσολογέω -ῶ / διττολογέω -ῶ; μονόβιβλος, ὁ / μονόβιβλον, τό). The grammatical category of the headword is registered in the **Grammar** section (Fig. 2) under the labels of Noun, Verb, Adjective, Adverb, Participle, and Verbal Adjective. Further distinctions or tagging might be possible, allowing search options for substantivized items, prepositional phrases (ἀπό κοινοῦ), or lexical phrases/multi-word terms (κοινός τόπος). Adverbs and participles are not subordinated under Adjective and Verb, as they prove crucial for ancient Greek scholarly terminology. A distinct field exists for recording the corresponding **Latin** term (Fig. 3) as attested in Latin grammar and relevant treatises. We shall complete this task at a later stage. It will be of great value for tracing the continuity between the ancient and modern grammatical/scholarly terminology, as well as the connection between the two classical languages in this domain: for example, the direction of borrowing for pairs, such as ἄκλιτος/ἄπτωτος > indeclinabilis/aptotus and glossema > γλώσσημα. Finally, under the **Dictionaries** field (Fig. 4), the presence of a term in the primary general dictionaries of Ancient Greek (LSJ and LSJSuppl., DGE, GI, GE) is recorded, either if it is listed as a separate headword or within a relevant entry, as regularly in the case of adverbs. The inclusion of dictionaries in this list was based on their lasting influence (LSJ) and being newly published (GE). The list of reference dictionaries might expand to include new relevant publications.

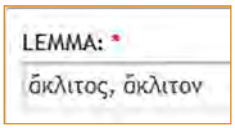


Fig. 1

Fig. 2

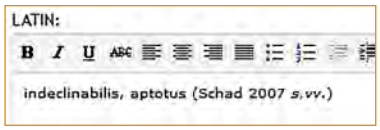


Fig. 3

Fig. 4

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

2.3.2 The Main Lemmatical Structure

The interdisciplinary character of ancient Greek scholarship manifests itself in the **Field** domain, where we allocate one or more scholarly fields to each term based on its meaning and specific usage. The key fields are Grammar, Language, Literary Criticism, Meter, Philology, and Rhetoric. Sometimes the attribution of a term to a single Field may be conventional, given the blurring borders between them. Quite typical are the cases of terms related to prosody, as they can refer either to Grammar and Meter (Fig. 5) or grammatical terms frequently found within a rhetorical or literary context and, hence, correspondingly marked (Fig. 6).

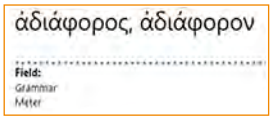


Fig. 5

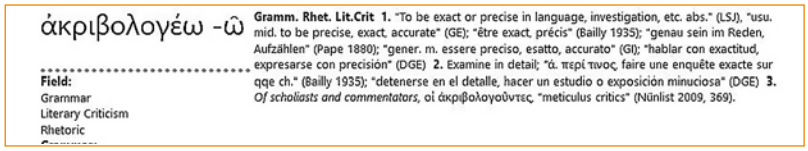


Fig. 6

Vocabula Grammatica is not a monosemous terminological guide but a linguistic dictionary that foregrounds the polysemy of the philological terms by tracing their use in various contexts. This character reflects in the **Translation** section. Here, the entries thoroughly display their respective meanings and the particular use of individual forms or collocations based on a meticulous (sub-)categorization of the available evidence. Hence, we opt for a multi-layered translation, where a term is being defined, first, according to its field of reference (I, II, III), secondly, to its meaning or sub-meanings (1, 2, 3), and finally, according to its specific usages (a, b, c / i, ii, iii) (Fig. 7, for ἀναδιπλόω).

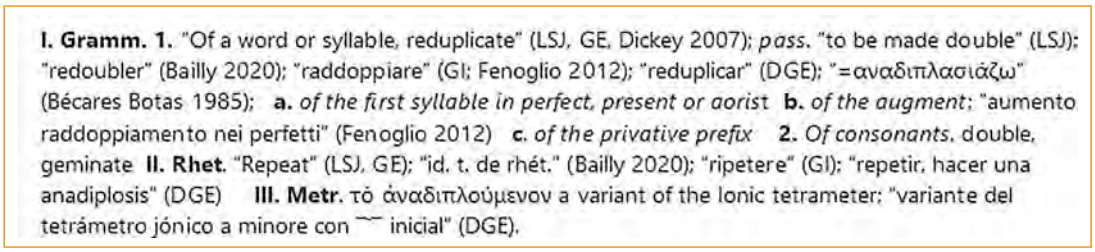


Fig. 7

Due to its broad chronological coverage, the dictionary discourages a saliency-based classification of meanings in favor of a chronological one (see also in **Sources**). The saliency of a meaning manifests itself basically in the number of relevant quotations. While the headword and the sources are in Ancient Greek, English is the principal working language; additionally, we record translations and interpretations in French, German, Italian, and Spanish. Under the relevant meaning, we cite available translations in the basic general dictionaries of Ancient Greek (LSJ, GE, Bailly, Pape, GI, DGE) along with other specialized glossaries, studies, or (translated) editions (see **2.1**). As much as this practice seems to correspond to a pre-final stage of compilation, it constitutes a distinctive characteristic of this dictionary, bringing together direct with indirect sources. There are entries where we acquire the translation exclusively from existing definitions and others where we supplement it. Besides, there are entries where we provide a wholly new translation, when either the

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

term itself or its grammatical sense does not testify to the relevant indirect sources. In any case, existing definitions are not merely juxtaposed but critically embedded in the arrangement of meanings and usages. This multilingual and polyphonic character is further enhanced by providing the Latin equivalent. The **Translation** domain sets a series of issues of concern for the compiler of a specialized dictionary. We are here confined to mention two of them. A fascinating interplay occurs between general and special meaning as we come upon familiar words supplied by new meaning(s) within their specific usage (for example, κρᾶσις, ἀόριστος, ἐνέργεια, πρόσωπον, ψιλός), or – vice versa – words and meanings that move from the special to the general vocabulary (for example, μεταφορά, ἀριθμός, βραχυσύλλαβος). Moreover, we attempt to introduce the terms as they were used and understood by ancient authors (see Dickey 2007), allowing even for consistencies and inconsistencies to emerge; thus, we try to avoid anachronisms that may further complicate the already unstable landscape of ancient Greek scholarship (see below **3**).

The **Sources** domain immediately follows the **Translation**. By definition, historical dictionaries, such as the *Vocabula Grammatica*, rely much upon:

> [...] snippets of text from cited sources. The aim is not only to show the word in context but also to show that it exists, precisely, at a particular date and in a particular source; the citation material is the verifiable documentary evidence on which the entry is built. (Hanks/de Schryver 2015, p. 7)

Initially, the compiler scans the whole kaleidoscope of sources in ancient and medieval Greek texts by applying and using the *Thesaurus Linguae Graecae* (TLG) – the online corpus of Ancient Greek and Byzantine literature – as well as by consulting critical editions not yet included in the TLG, mainly based on our canon of primary sources (see **2.2**). The survey, study, and analysis of sources allow the compiler to arrange them both in the horizontal (chronological, historical) and the vertical (interdisciplinary, semantic, interpretive, or explanatory) axis. This indexation facilitates the parallel study of the development of the meaning or meanings of each entry term. As the *Vocabula Grammatica* dictionary does not aim at exhaustiveness but rather at representativeness and relevance, the quantity of the citations depends on these criteria (cf. Ashdowne 2016, p. 354). Where possible, especially in the case of terms with a small number of attestations in Ancient Greek and Byzantine literature, the indexation is exhaustive (ἀοριστώδως/ἀοριστωδῶς). Yet, we record the presence of each term in the relevant literary tradition throughout the whole period of its attestation. Consequently, there are terms whose attestation at the corpus ranges from an *hapax legomenon* or a few dozen to hundreds or thousands of references, an element captured in the **Sources** domain.

At the microstructure level (see Fig. 8, for ἀναδιπλόω), the sources are indexed chronologically within each (sub)category of field, meaning, and usage – definitely according to the first attestation of the term – to facilitate a fair general overview of its semantic development. So, we also provide not mere citations but extensive quotations from sources. Each lemma is simultaneously composed by studying the direct (the texts) and the indirect (the dictionaries) sources, though first based on the chronological and then its generic and semantic arrangement. The study of the whole corpus allows the compiler to gain overall supervision of the literary tradition through its intermittent transmission or even copying in the scholarship continuum. At the end of the Sources section, we array various undated scholia according to the author's era and the chronology of the original treatise on which they comment. All three domains, namely the **Field**, **Translation**, and **Sources**, are strongly interrelated and inherently contextual.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

I. 1. a. 1. Trypho *Path.* 1.21 ἄρσις δέ ἐστιν ἀποβολή κατ' ἀρχήν τῆς **ἀναδεδιπλωμένης** συλλαβῆς, οἷον βεβλῆσθαι βλῆσθαι, δεδέχθαι δέχθαι, βεβλημένος βλήμενος, δεδεγμένος δέγμενος. Also 2.5, 3.5. Cf. Rhacend. *Synop.* 18.566.17. 2. Heracl. fr. 28.28 : Eust. *Comm. Od.* 1722.50 (12.295) Ἴωνες, ὅτ' ἂν **ἀναδιπλῶσι** ῥήματα, τὸ αὐτὸ ἀρκτικὸν ποιοῦνται πρώτης καὶ δευτέρας συλλαβῆς· λαβέσθαι λελαβέσθαι, κάμω κεκάμω, πιθέσθαι πεπιθέσθαι, δάσκω διδάσκω, τύσκω τιτύσκω, χωρὶς εἰ μή μέλλοι δύο δασέα κείσεσθαι παράλληλα. 3. Heracl. fr. 28.44 : Eust. *Comm. Od.* 1722.62 (12.295) ὅτε δὲ γράμματι ἐνθεωρεῖται ἐνὶ ἡ ἀναδίπλωσις, τότε ἰσοσυλλαβεῖ τῷ **ἀναδεδιπλωμένῳ** τὸ ὁλόκληρον, οἷον μένω μίμνω, γένω γίγνω. 4. Heracl. fr. 39.21 : Eust. *Comm. Od.* 1613.224 (9.19) Βοιωτοὶ τὸ ἦτα παρατέλευτον ὂν τῆς πρώτης τῶν εἰς μι μετατιθέντες εἰς τὴν ει δίφθογγον πολλάκις καὶ **ἀναδιπλοῦσιν** οἷον πεφίλειμι νενόειμι. 5. Phryn. *SP* 32.12 ἀληλιμμένον: **ἀναδιπλοῦντες** λέγουσιν (sc. οἱ Ἀττικοί) ἀντὶ τοῦ ἠλειμμένον. οὕτω καὶ τὸ κατορώρυκτο καὶ τὸ κατορωρυγμένον. Cf. An. *Syn.* α 971.1; Phot. *Lex.* α 939.1; An. *Seg.* α 74. 6. Oros. 6a.13 οὐ μὴν ἐν ἅπασί γε τοῖς συνθέτοις τὰς προθέσεις οἱ Ἀττικοὶ φυλάττουσι, ἀλλὰ εἰσιν ἀνώμαλοι καὶ ἐν τούτῳ. ἐπεὶ οὖν πολλῶν **ἀναδιπλοῦσι** τὰς προθέσεις, λέγουσι γοῦν καὶ 'δεδιακόνηκα' (Dem. 51,7) καὶ 'δεδιῴκηκα' καὶ ἄλλα τοιαῦτα. Cf. L. *Zon.* α 214.2. 7. Oros. 6a.21 καίτοι οἱ γραμματικοί φασιν „αἱ προθέσεις οὐκ **ἀναδιπλοῦνται**". **b.** Eust. *Comm. Od.* 1.6.11 τὸ δὲ πλάγχθη ἀναύξητον ὂν, 'Ιαστὶ

Fig. 8

2.3.3 Cross-reference structure

As has been pointed out:

> Cross-reference structure is a lexicographical term for the arrangement of those explicit and implicit indicators that direct the user within the dictionary for additional or supplementary information over and above that already found at the first lookup. [...] Used correctly, cross-references play a major role in the dictionary, in that they serve to unify and amplify the information provided, thus giving the user a more comprehensive view. (Bergenholtz/Tarp (eds.) 1995, p. 16)

That's why the dictionary includes the domains of **Bibliography**, **Notes**, and **Related Terms**. These three sections connect us to the microstructure of each entry (see **2.3**), the macrostructure (namely the overall wordlist), and the outside matter of the dictionary. In the **Bibliography** section, we provide bibliographic references used for the compilation of each entry, thus further enhancing the comprehension and documentation of every term. Within the **Notes**, users may find various kinds of remarks: a) bibliographic information, such as encyclopedic type of information, notes concerning textual criticism, enlightening information in reference books and articles; b) information regarding each specific lemma, such as "Term not attested in dictionaries" "Exhaustive indexation" or "Hapax found" in sources, variant or dialectic forms of the term itself, comments on probable erroneous explanations, misconceptions or misinterpretations in other dictionaries and reference works, or allusions to the general vocabulary (cf. ἀντονομασία/ἀντωνομασία, ἀντονομία, ἀοριστώδης, ἀστερίσκος, βραχυσύλλαβος, ὀλιγοσύλλαβος, συλλαβίζω). In the **Related Terms**, each entry transcends the microstructure and communicates again with the macrostructure of the dictionary. The Related terms are selected according to their morpho-semantic and conceptual relation with the listed lemma (Fig. 9, for ἄκλιτος). Therefore, 'families' of either derivative, compound, or semantically related words may emerge.

Related terms:
ἀκλίτως
ἀκίνητος, ἀκίνητον
ἄπτωτος, ἄπτωτον
κλίσις, ἡ
κλίνω
μονόπτωτος, μονόπτωτον

Fig. 9

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

3. Philological issues

To this day, the lexicography of the ancient Greek grammatical terms presents certain complications and problems; some are mentioned immediately below.

a) The sources. From the 3rd century BC to the 3rd century AD, few works of Greek grammarians and philologists survived to be transmitted directly. Most grammatical and philological treatises have since been lost or transmitted only indirectly through references or quotations in later – mostly Byzantine – commentaries, epitomes, and dictionaries. Yet, some collections exist with fragments of those now-lost written works preserved in later sources that partially compensate for their loss. However, since the grammatical theory and terminology evolved considerably between the 3rd century BC and the 15th century AD, we cannot exclude the possibility that later sources might have rephrased the wording of earlier grammarians, using and applying the terminology of their times. Because of these substantial gaps in the tradition and the availability of indirect rather than direct sources for the works of many early grammarians, it is often difficult to identify the first appearance of many terms; it is also difficult to reconstruct their semantic development and their relation to other grammatical terms.

b) The editions. Still, there are no reliable editions for several texts belonging to the grammatical and rhetorical corpora (especially for many Byzantine commentaries, vocabularies, and treatises); in such cases, it is necessary to resort to old editions of the 19th century. Therefore, publications like Erbse's edition of the *scholia graeca* on the Iliad and Pontani's ongoing analogous edition on the Odyssey, Van der Valk's edition of Eustathius of Thessalonica's Commentary on the Iliad, or Koster's et al. multi-volume scholia on Aristophanes are more than welcome.

c) Fluidity in terminology. As Eleanor Dickey has put it:

> [...] there is a certain fluidity in Greek technical terminology, so that the same word can have a number of different uses in different passages. Often these differences are the result of the evolution of grammatical theory during the thousand or so years in which ancient scholarship developed. [...] Sometimes, however, a single word can have a variety of uses even within one grammatical treatise; for example, Dionysius Thrax uses ἀόριστος both to mean "aorist tense" and to mean "indefinite." (Dickey 2007, p. 124)

The same is true to an even greater extent of the terminology in the disciplines of rhetoric and literary criticism.

4. Conclusions – further perspectives

Compiling a dictionary marked by such a vast generic and chronological range is not an easy task: it demands steady, long-term, and meticulous teamwork, which is fortunately facilitated nowadays by the systematic use of digital tools concerning either the search or the management of the content (digital corpora and dictionaries on the one hand; CMS on the other). Aiming at exploiting further the possibilities offered by more sophisticated content management systems, our plans in the immediate future include:

– Making more subtle distinctions at the level of content management, thus allowing for more detailed, analytical, and informative search options: for instance, dating the authors or searching into different authors and texts.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

- Elaborating on the cross-reference structure in three directions: a) to other articles of the inner matter (Related terms); b) to the outer matter (abbreviations, bibliographical references, list of editions used); c) aspiringly, to the outside matter of the dictionary (external literature and relevant digital tools, that is other dictionaries).
- Developing the Related terms domain, so as to achieve a more detailed classification based on linguistic (morpho-semantic) and conceptual criteria, namely synonyms, opposites, and concept maps.

Most likely, the completion of the *Vocabula Grammatica* dictionary will make salient the systematic character of the ancient Greek scholarship along with the ideological and cultural parameters that make it an autonomous scientific field, as well as its interconnection with other disciplines of antiquity, though without ignoring the methodology and issues of modern literary criticism. So, it will undoubtedly contribute to the renewal and enrichment of the general dictionaries of the ancient Greek language within the emerging and promising field of Digital Humanities, offering a kind of a digital Ariadne's String for navigating into the labyrinth of ancient Greek scholarship.

References

Anderson, R. D. (2000): Glossary of Greek rhetorical terms connected to methods of argumentation, figures and tropes from Anaximenes to Quintilian. (= Contributions to Biblical Exegesis and Theology 24). Leuven.

Ashdowne, R. (2016): Dictionaries of dead languages. In: Durkin, P. (ed.): The Oxford handbook of lexicography. (= Oxford Handbooks in Linguistics). Oxford, pp. 350–366.

Bailly, M. A. (1935, 2020): Dictionnaire Grec-Français. Nouvelle édition revue et corrigée, dite Bailly 2020-Hugo Chávez, établie sous la direction de G. Gréco et al. Paris.

Bécares Botas, V. (1985): Diccionario de terminología gramatical griega. (= Acta Salmanticensia, artes dicendi. Fuentes para la lingüística retórica y poética clásicas III). Salamanca.

Bergenholtz, H./Tarp, S. (eds.) (1995): Manual of specialised lexicography: The preparation of specialised dictionaries. (= Benjamins Translation Library 12). Amsterdam/Philadelphia.

De Jonge, C. C. (2008): Between grammar and rhetoric: Dionysius of Halicarnassus on language, linguistics and literature. (= Mnemosyne Supplements. Monographs on Greek and Roman Language and Literature 301). Leiden/Boston.

DGE = Adrados, F. R. et al. (1989–2009): Diccionario Griego-Español. Vols. I–VII (α - ἕξαυος). Madrid.

Dickey, E. (2007): Ancient Greek scholarship: a guide to finding, Reading, and understanding scholia, commentaries, lexica, and grammatical treatises, from their beginnings to the Byzantine period. (= American Philological Association, Classical Resources Series). Oxford.

Erbse, H. (ed.) (1983, 1988): Scholia graeca in Homeri Iliadem (scholia vetera). Vols. VI–VII (Indices). Berlin.

Ernesti, C. T. (1795): Lexicon technologiae Graecorum rhetoricae. Leipzig.

Fenoglio, S. (2012): Eustazio di Tessalonica Commentari all'Odissea: glossario dei termini grammaticali. (= Hellenica. Testi e strumenti di letteratura greca antica, medievale e umanistica 43). Alessandria.

GE = Montanari, Fr. (2015): The Brill dictionary of Ancient Greek. Editors of the English Edition M. Goh/Ch. Schroeder. Leiden.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

GI = Montanari, Fr. et al. (1995): Vocabolario della lingua greca. 3rd ed. 2004. Torino.

Hanks, P./de Schryver, G.-M. (eds.) (2015): International handbook of modern lexis and lexicography. Berlin/Heidelberg.

Keizer, H. M. et al. (1995): Indices in Eustathii archiepiscopi Thessalonicensis Commentarios ad Homeri Iliadem pertinentes ad fidem codicis Laurentiani editos a Marchino van der Valk. Leiden/New York/Köln.

Lausberg, H. (1998): Handbook of literary rhetoric: a foundation for literary study. Translated from German. Edited by D. E. Orton/R. Dean Anderson. Leiden.

LSJ/LSJSuppl. = Liddell, H. G./Scott, R./Jones, H. S. et al. (1940, 1996): A Greek-English lexicon. 9th edition (1940) with a revised supplement (1996). Oxford.

Martin, J. (1974): Antike Rhetorik: Technik und Methode. (= Handbuch der Altertumswissenschaft 2.3). München.

Matthaios, S. (2014): Philological-grammatical tradition in ancient linguistics. In: Giannakis, G. K. et al. (eds.): Encyclopedia of Ancient Greek Language and Linguistics (EAGLL). 3 volumes. Leiden/Boston, vol. 3, pp. 63–76.

Matthaios, S. (2015): Greek scholarship in the imperial era and late antiquity. In: Montanari, F./Matthaios, S./Rengakos, A. (eds.): op. cit., vol. 1, pp. 184–296.

Matthaios, S./Montanari, F./Rengakos, A. (eds.) (2011): Ancient scholarship and grammar: archetypes, concepts and contexts. (= Trends in Classics, Supplementary Volumes 8). Berlin/New York.

Meijering, R. (1987): Literary and rhetorical theories in Greek scholia. Groningen.

Montanari, F. (2011): Ancient scholarship and classical studies. In: Matthaios, S./Montanari, F./Rengakos, A. (eds.): op. cit., pp. 11–26.

Montanari, F. (ed.) (2020): History of ancient Greek scholarship: from the beginnings to the end of the Byzantine Age. Leiden.

Montanari, F./Matthaios, S./Rengakos, A. (eds.) (2015): Brill's companion to ancient Greek scholarship. 2 volumes. Leiden.

Novokhatko, A. (2015): Greek Scholarship from its Beginnings to Alexandria. In: Montanari, F./Matthaios, S./Rengakos, A. (eds.): op. cit., vol. 1, pp. 3–59.

Nünlist, R. (2009): The ancient critic at work: terms and concepts of literary criticism in Greek scholia. Cambridge.

Pape, W. (1880): Handwörterbuch der Griechischen Sprache. Griechisch – Deutsches Handwörterbuch Bd. I (A–K); Bd. II (Λ–Ω). 3. Auflage. Berlin.

Robins, R. H. (1993): The Byzantine grammarians: their place in history. (= Trends in Linguistics, Studies and Monographs 70). Berlin/New York.

Sluiter, I. (1990): Ancient grammar in context: contributions to the study of ancient linguistic thought. Ph.D. thesis. Amsterdam.

Tikkanen Westin, K. (2018): Chrysoloras' Erotemata, and the evolution of grammatical description. In: Beiträge zur Geschichte der Sprachwissenschaft 28 (1), pp. 33–56.

Urrea Méndez, J. (2003): El léxico métrico de Hefestión. (= Classical and Byzantine Monographs LIV). Amsterdam.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Contact information

Maria Arapopoulou
Centre for the Greek Language, Thessaloniki, Greece
marapopoulou@gmail.com

Georgios Kalafikis
Centre for the Greek Language, Thessaloniki, Greece
gkalafikis@gmail.com

Dimitra Karamitsou
Centre for the Greek Language, Thessaloniki, Greece
dimika91@gmail.com

Efstratios Sarischoulis
Centre for the Greek Language, Thessaloniki, Greece
efstratios.sarischoulis@gmail.com

Sotiris Tselikas
Centre for the Greek Language, Thessaloniki, Greece
stselik@gmail.com

Acknowledgements

We would like to thank all the colleagues who have contributed to the elaboration of this laborious task; special mention should be made to Anna Lichou, Agori Antonopoulou, and Dimitris Manios.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Anaïs Chambat

LA LIGNÉE «CAPURON-NYSTEN-LITTRÉ» ENTRE RUPTURES ET CONTINUITÉS DOCTRINALES

Abstract This article aims to show the influence of doctrines in the medical lexicographers choices, with the Capuron-Nysten-Littré lineage as a case study. Indeed, the *Dictionnaire de médecine* has been crossed by several schools of thought such as spiritualism and positivism. While lexical continuity may seem self-evident due to the nature of the work, thus reducing the reprint to a simple lexical increase, this process introduces neologisms and deletions, all can be considered in their effects by using text statistics and factorial analysis.

Keywords History of medical science dictionaries; 19th–20th centuries; medical lexicographers; spiritualism; positivism; text statistics; Joseph Capuron; Pierre-Hubert Nysten; Émile Littré; Charles Robin; Augustin Gilbert; Medica

1. Introduction

La multiplication des dictionnaires des sciences médicales au XIXe siècle, tout comme l'accroissement de ces ouvrages au fil de leurs rééditions, font d'eux les témoins des progrès scientifiques accomplis. Toutefois, les choix lexicaux peuvent également refléter des choix doctrinaux. L'objectif de cette recherche est de fournir un premier aperçu de ces deux aspects à partir de l'histoire de la lignée « Capuron-Nysten-Littré », depuis la parution en 1806 du *Nouveau dictionnaire de médecine, de chirurgie, de physique, de chimie et d'histoire naturelle* de Joseph Capuron, jusqu'à la 21e et dernière édition du *Dictionnaire de médecine, de chirurgie, de pharmacie et des sciences qui s'y rapportent*, d'Emile Littré et d'Augustin Gilbert. À partir des données lexicographiques collectées dans le cadre du projet CollEx-Persée « Métadictionnaire médical multilingue »[1] de la bibliothèque numérique française Medica,[2] nous avons constitué un corpus de huit dictionnaires: Capuron (1806), Nysten (1833) – en réalité Briand/Bricheteau/Henry (1833), Littré/Robin (1855, 1865 et 1873), Littré/Gilbert (1908) ainsi que deux autres utilisés à des fins comparatives, Lavoisien (1793) et Bégin et al. (1823). Il s'agira notamment d'évaluer le degré de parenté entre eux et de mesurer l'influence des doctrines philosophiques, scientifiques et morales de leurs auteurs. Notre étude se fondera d'une part sur une analyse qualitative des préfaces, des comptes rendus bibliographiques ainsi que des témoignages recueillis dans divers journaux médicaux. Par ailleurs, elle s'appuiera sur des données quantitatives inédites, telles que des statistiques sur l'évolution du nombre de vedettes ou encore la nomenclature des lemmes supprimés ou ajoutés; une analyse de séquences et une analyse en correspondances multiples permettront de modéliser l'évolution de leur répartition sur l'ensemble des huit dictionnaires.

[1] Pour en savoir plus sur le projet: voir la page dédiée sur le site de CollEx-Persée.

[2] Lien vers les « Dictionnaires de Medica ».

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

2. Aux origines de la lignée « Capuron-Nysten-Littré »

« Rien n'est plus important pour ceux qui cultivent une science, que d'en bien connaître la langue » (Savary 1810, p. 393). Publié en 1764, le *Dictionnaire portatif* du chirurgien Jean-François Lavoisien a pour ambition de former les étudiants (Lavoisien 1793, p. 3). L'une des particularités de cet ouvrage est notamment la présence d'un vocabulaire des termes grecs et latins fournissant « un modèle à ceux qui sont venus après » (Dechambre 1869, p. 106). Il s'agissait de disposer d'un volume unique, de petit format et d'usage facile contenant l'essentiel sur « l'art de guérir ». Véritable pont entre la fin du XVIIIe et le milieu du XIXe siècle, le *Nouveau dictionnaire de médecine* de l'obstétricien catholique Joseph Capuron est d'un grand intérêt lexical. Basé sur la nomenclature et le contenu du Lavoisien (Chaumeton 1814, p. 271), il comprend deux glossaires grec et latin. Une attention particulière est portée à l'existence de variations orthographiques et à l'origine des mots.[3] Las de « l'insuffisance des anciens vocabulaires », et conscient de la nécessité « d'en composer de nouveaux qui soient à la hauteur des connaissances actuelles » (Capuron 1806, p. V), il y consacre les termes adoptés par les savants.

Face à l'engouement du public, il collabore avec Pierre-Hubert Nysten pour la deuxième édition parue en 1810. Grâce à la concision de ses définitions et à sa complétude, la troisième édition de 1814, entièrement refondue par Nysten seul, devient incontournable. Après sa mort, en 1818, Isidore Bricheteau, Étienne Henry et Joseph Briand poursuivent son œuvre. Ils s'engagent notamment à « n'omettre aucun mot utile [et] à laisser dans l'oubli les mots bizarres dont un ridicule néologisme embarrasse chaque jour le langage médical » (Briand/Bricheteau/Henry 1833, p. V). En 1845, l'éditeur médical Baillière en rachète les droits. En 1855, Émile Littré et Charles Robin le refondent complètement au motif que les avancées scientifiques le nécessitent.[4] L'ouvrage est supposé s'inscrire dans la tradition de Nysten dont il conserve le patronage, mais il a une autre ampleur. *De facto*, le dictionnaire élémentaire se transforme en encyclopédie. « De nombreuses figures gravées avec exactitude, et intercalées dans le texte » (Littré/Robin 1855, p. VI) augmentent encore son utilité de même que ses six glossaires (latin, grec, allemand, anglais, italien et espagnol) qui contiennent les principaux *items* de la langue médicale.

3. D'une école de pensée à l'autre

Cependant, il ne s'agit pas du seul tournant majeur opéré à partir de l'édition de 1855. Elle est en effet marquée par la disparition du point de vue « spiritualiste » (Dechambre 1879, p. 853), chrétien, moniste et universaliste adopté par Capuron puis Nysten. D'après ce courant de pensée, la métaphysique permet d'apprendre le fonctionnement de la nature,[5] au même titre que la science, soit un holisme transdisciplinaire entendu comme base de la connaissance scientifique.[6]

---

3 « Lorsque l'étymologie est connue, elle doit servir de règle à cet égard, à moins qu'un usage très ancien et très général n'ait prévalu ». (Savary 1810, p. 396).

4 « Le progrès des sciences médicales ne permettant pas qu'on se contentât d'une réimpression, il fallait en venir à un remaniement ». (Littré/Robin 1855, p. V).

5 « [C'est] l'esprit qui donne la clé de la nature ». (Gouhier 1999, p. 26).

6 « Biran, Ravaisson, Lachelier, Bergson… vues de haut, leurs œuvres tracent une même ligne qui symbolise le mouvement du spiritualisme en France au XIXe siècle ». (*Ibid.*, p. 20).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Elle repose sur le principe fondamental que « l'âme est formée du corps » (Giraud 1862, p. 3), ce qui permet l'établissement d'un « ensemble cohérent et logique », d'une « unité réelle et profonde dans l'œuvre entière »,[7] la philosophie positiviste est ainsi au cœur des choix lexicographiques de Littré/Robin. Suivant les préceptes d'Auguste Comte, la science doit éprouver ses hypothèses au contact de l'expérience qui, elle seule, les validera. L'esprit est sans cesse rappelé à la réalité: il faut renoncer aux idées et se river aux faits. Dans le domaine de la biologie, il ne peut être de fonction qui ne soit liée à un organe ; inversement, tout substrat organique assure un rôle (Dagognet 1997, p. 937). La lecture des préfaces et des comptes rendus bibliographiques permet de dresser une liste non exhaustive des termes[8] qui mettent en relief les opinions philosophiques, scientifiques et morales de leurs auteurs, dont voici quelques exemples:

- **ÂME:** Capuron définit l'âme comme un « souffle, [un] principe de vie » (Capuron 1806, p. 16), Nysten tel le « principe des facultés intellectuelles et affections morales » (Briand/Bricheteau/Henry 1833, p. 68). L'influence des Saintes Écritures et de la pensée de Saint-Thomas d'Aquin,[9] elle-même héritière de la philosophie antique, est palpable. Dans le Littré/Robin, l'âme devient « l'ensemble des fonctions du cerveau et de la moelle épinière » (Littré/Robin 1855, p. 57). Entre 1865 et 1908, un avertissement est intégré au texte: « cette définition résulte du dogme scientifique actuel, qui n'admet ni propriété ou force sans matière […] tout en déclarant ignorer ce que c'est en soi que force et matière ». (Littré/Robin 1865, p. 55; 1873, p. 54). Les auteurs admettent ici l'existence de réalités dont ils ne connaissent pas l'essence, et qui dépassent leur vision du monde.
- **HOMME:** Si Capuron définit l'homme comme « le plus parfait des êtres organisés » (Capuron 1806, p. 165), Nysten fait référence à son intelligence, à son aptitude à avoir des idées, à les classer et à les exprimer, à sa mémoire, à son jugement ou encore à son imagination (Briand/Bricheteau/Henry 1833, p. 484). Littré/Robin classent l'homme dans l'ordre animal, et consacrent de longs développements à la notion de race et de variété dans le genre humain (Littré/Robin 1855, p. 634–637). En 1908, afin de faire taire toute polémique, Gilbert y rajoute une introduction restrictive: « L'homme, considéré au point de vue purement zoologique, peut être défini un animal mammifère de l'ordre des primates et de la famille des bimanes, caractérisé taxonomiquement par une peau » (Littré/Gilbert 1908, p. 801).
- **MORT:** Suivant Capuron, « la mort est la séparation de l'âme d'avec le corps, qui n'est plus alors qu'une masse inerte, froide et insensible, un cadavre » (Capuron 1806, p. 220). Dans le système positiviste, la vie n'étant qu'une « manifestation des propriétés inhérentes et spéciales à la matière organisée » (Littré/Robin 1855, p. 1341), la mort est la cessation définitive de cette manifestation. Il serait donc toujours possible de trouver une

[7] « Dans un Dictionnaire, qui, assujetti à l'ordre alphabétique, ne procède que par fragments isolés, il est difficile, mais fort important d'avoir une philosophie qui, par un lien secret, réunisse les parties éparses [et] donne l'enchaînement du savoir humain ». (Littré/Robin 1855, p. VI).

[8] Nous avons aussi pu relever: *animisme*, *cause*, *corps*, *entendement*, *esprit*, *être*, *fonction*, *force*, *forme*, *humanité*, *idée*, *induction*, *inertie*, *innervation*, *insénescence*, *instinct*, *irréductibilité*, *localisation*, *logique*, *loi*, *maladie*, *matérialisme*, *matière*, *médecin*, *médicament*, *mésologie*, *métaphysique médicale*, *moral*, *nature*, *nosologie*, *pathologie*, *pensée*, *phénomène*, *positive (philosophie)*, *praticien*, *pratique*, *propriété*, *rationalisme*, *résultat*, *science*, *sens*, *sensation*, *sensibilité*, *sentiment*, *songe*, *spécialiste*, *spéculative (médecine)*, *spiritualisme*, *subjectif*, *syphilis*, *thérapeutique*, *végétalité*, *vie*, *virulence*, *vitalisme* et *vivisection*. (Voir notamment Guardia 1865, p. 203–206; 218–220).

[9] « L'homme ne doit pas être regardé comme une âme ayant un corps à son service (*anima utens corpore*) ; […] l'homme est un composé d'âme et de corps, et ces deux composants de sa nature forment une seule substance, un être unique, qui est le sujet de la connaissance sensible, comme de toutes les autres opérations humaines ». (Moreau 1976, p. 6).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

cause physique au décès d'un individu. Cependant, comme l'observe Giraud (1862, p. 10), il y a des maladies qui ne laissent aucune trace dans les tissus organiques, et qui pourtant, entraînent la mort.

– **RAISON:** À partir de Littré/Robin, il ne s'agit plus de « distinguer le bien du mal » (Briand/Bricheteau/Henry 1833, p. 786), mais de « démontrer le vrai ». Par ailleurs, « on observe chez beaucoup d'animaux une appréciation judicieuse des circonstances » (Littré/Robin 1855, p. 1055), la raison ne serait donc pas l'apanage de l'homme.

Littré et Robin paraissent donc comme les « adeptes d'un rationalisme scientifique, voire d'un scientisme » (Sournia 1981, p. 230). On pourrait penser que cette philosophie n'atteint qu'un public limité. Or, grâce à cet instrument de travail indispensable,[10] elle touche un plus vaste public (Sournia 1981, p. 231). À la sortie de la douzième édition du *Dictionnaire* en 1865, les éditeurs « ne présentent plus l'ouvrage comme étant l'œuvre de Nysten, ni même comme étant fait *d'après le plan de Nysten*, mais bien *d'après le plan suivi par Nysten.* [...] Nysten avait lui-même suivi et augmenté le plan de Capuron » (Littré/Robin 1873, p. VIII). Il est fait référence ici à l'enrichissement cumulatif de la nomenclature du *Dictionnaire*. Encouragée par le parti clérical, la veuve de Nysten assigne les éditeurs devant le tribunal de la Seine. Le 18 mai 1865, en première instance, les libraires éditeurs Baillière sont reconnus comme propriétaires du livre. La veuve fait appel. Elle leur reproche en effet d'avoir placé sous le nom de son mari vitaliste et spiritualiste une philosophie toute autre. Le 27 février 1866, les éditeurs sont condamnés à lui verser des dommages et intérêts (Dechambre 1879, p. 854). Littré meurt le 2 juin 1881. A son tour, sa veuve exige que certaines définitions du Littré/Robin, qu'elle juge regrettables, soient modifiées (Sournia 1981, p. 233). Le *Dictionnaire* est censuré: dans l'édition de 1908, toute allusion positiviste est supprimée de l'article « âme ».

4. « L'effort inventif de toute une époque »[11]

Le *Dictionnaire des termes de médecine, chirurgie, art vétérinaire, pharmacie, histoire naturelle, botanique, physique, chimie, etc.* dirigé par Louis-Jacques Bégin est publié en 1823, soit la même année que la quatrième édition du *Dictionnaire*. Les auteurs ont pour ambition de produire un lexique « aussi court et complet que possible » (Bégin 1823, p. VI). Rassembler 11 215 entrées en un peu moins de 600 pages n'est pas une mince affaire. Ils se sont concentrés sur les « mots qui revenaient le plus souvent dans les livres et les cours de médecine » (Bégin 1823, p. VI), mais dont la signification n'était pas nécessairement fixée.

Cette originalité suscite des critiques. Un contemporain note même: « il est surchargé de mots hétéroclites, de définitions ridicules, qui en font un livre véritablement burlesque » (Duryer 1824, p. 2). De nouveaux mots, rencontrés dans aucune autre nomenclature étudiée, sont introduits,[12] parmi lesquels au moins une cinquantaine ont trait à la divination;[13] enfin,

[10] « Il forme une encyclopédie présentant, à cause de la rapidité avec laquelle les éditions se succèdent, un tableau exact de la science. C'est ainsi qu'il peut servir à la fois de *vademecum* au praticien et au savant, de mémorial au maître et à l'élève, de guide à tous ceux qui désirent, au milieu de la diffusion actuelle des sciences, ne pas rester étrangers à ce mouvement ». (Littré/Robin 1865, p. V).

[11] Titre extrait de la préface du Littré-Gilbert (1908, p. V).

[12] Sur les 2 253 mots concernés, soit 20% de l'ouvrage, citons entre autres : *abrodiète*, *adelphixie*, *aesthème*, *agnoie*, *algola*, *ampharistère*, *anarrhégnyme*, *anastaechéiose*, *anisotaque*, *apenthysmène*, *aphepsème*, *apomathisie*, *begma*, *cagastrum*, *coptarion* et *scodeghino*. (Duryer 1824, p. 2).

[13] Nous avons effectué cette estimation à partir des entrées ayant pour suffixe « -mancie ».

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

de nombreux sont empruntés aux dictionnaires de langue française.[14] Néanmoins, près de 21% des lemmes de cet ouvrage sont communs aux huit dictionnaires du corpus, et 60% sont également présents dans les Nysten-Littré. L'édition du Nysten de 1833 a suivi cette voie en introduisant près de 9% de néologismes. Un critique note d'ailleurs à cet égard : « ce n'est pas à un dictionnaire de fermer la porte aux mots nouveaux ».[15] Il préconise de les considérer dans leur diversité, quitte à les supprimer lorsqu'ils seront tombés en désuétude. En choisir certains plutôt que d'autres, c'est afficher une préférence doctrinale, et par là-même, restreindre la portée scientifique et pratique de l'œuvre.[16]

À côté de ces ajouts parfois insolites, nous avons également constaté que les lexicographes ont supprimé pendant près de trente ans certains lemmes clés des rééditions de leurs dictionnaires. Par exemple, Bégin définit la « femme » telle la « femelle de l'homme » (Bégin 1823, p. 283). Elle n'apparaît qu'en 1865 en tant qu'individu à part entière (Littré/Robin 1865, p. 583). Jusqu'alors, elle était mentionnée en tant que « sage-femme », matrone ou accoucheuse.[17] D'un point de vue épistémologique, le cas de la « statistique » est intéressant, puisque c'est à cette période qu'elle s'institutionnalise en tant que discipline, méthode et pratique (Desrosières 1993). Deux articles distincts lui sont même consacrés au sein de l'*Encyclopédie m*éthodique (Maupertuis 1793, p. 612 ; Bricheteau 1830, p. 113). Définie par Capuron comme la « partie de l'économie politique qui a pour objet de fixer ou de faire connaître les richesses et les forces d'un état » (Capuron 1806, p. 320), elle est absente des éditions suivantes du *Dictionnaire*. Elle ne revient sous la forme de « statistique médicale » qu'en 1855 en tant que « détail de faits se rapportant aux morts, naissances, maladies, épidémies » (Littré/Robin 1855, p. 1179). Ce n'est qu'à partir de l'édition de 1865 que sa définition témoigne de la diversité des acceptions que recoupe la discipline (Littré/Robin 1865, p. 1428–1430). Amédée Dechambre précise d'ailleurs à cet égard que « nulle part peut-être le rôle social du médecin n'est plus manifeste, ni plus grand » (Dechambre 1864, p. XXXIII).

5. Quelques chiffres

D'un point de vue statistique, notre corpus est constitué par l'ensemble des vedettes, soit de 37 365 unités textuelles. Après le *Dictionnaire portatif* de Jean-François Lavoisien (1793), les entrées sont réparties en deux colonnes, ce qui permet d'augmenter le nombre de vedettes par page. L'ouvrage de Joseph Capuron (1806) a le plus petit format : son lexique ne s'étend en effet que sur 366 pages, le reste de l'ouvrage (28%) étant consacré aux synonymies. Pierre-Hubert Nysten l'enrichit considérablement: son contenu croît de 28% dès la troisième édition de 1814. Cette tendance est poursuivie par ses successeurs. L'édition de 1833 atteint les 956 pages, soit un accroissement volumétrique de 37%. La refonte opérée par Émile Littré

[14] Comme *acabit*, *accord*, *adroit*, *amour*, *bancroche*, *bout-en-train*, *bringue*, *cabriole*, *cadence*, *carnage*, *chant*, *chanter*, *chaudière*, *contrepied*, *danseur*, *destrier*, *ébat*, *réforme* ». (Duryer 1824, p. 2).

[15] « Il faut prendre et trier dans chaque époque les termes généralement adoptés, d'abord ceux qui, créés par nécessité, ont été nécessairement conservés, parce qu'ils n'étaient suppléés par aucun autre ; et puis ceux qui, rejetés plus tard et remplacés par d'autres dénominations plus précises, quoique sortis du langage médical actuel, sont utiles et même nécessaires à connaître pour apprécier les doctrines et pour comprendre les ouvrages de nos prédécesseurs ». (*Gazette médicale de Paris* 1833, p. 124).

[16] « Il perd ce caractère de généralité, d'impartialité au niveau des doctrines, pour ne refléter qu'un système, il lui devient impossible, sur une foule de points, de fournir des notions vraies, exactes, complètes ». (*Gazette médicale de Paris* 1872, p. 617)

[17] (Bégin 1823, p. 511); (Briand/Bricheteau/Henry 1833, p. 809); (Littré/Robin 1855, p. 1098; 1865, p. 1329; 1873, p. 1364); (Littré/Gilbert 1908, p. 1463).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

et Charles Robin pour la dixième édition de 1855, l'accroît de près de 500 pages et d'autant de figures intercalées dans le texte. Si l'édition de 1865 connaît une progression de 21%, il s'ensuit une hausse de 3% jusqu'en 1908 tandis que le nombre de figures connaît une croissance de 63%.

Afin de modéliser la répartition des entrées dédoublonnées sur l'ensemble des huit dictionnaires, nous avons choisi une analyse de séquences qui permet de considérer chaque élément comme une suite d'états, dans un espace fini de modalités. Elle vise à identifier dans la diversité d'un corpus, les régularités, les ressemblances, puis le plus souvent à construire des typologies de « séquences-types » (Robette 2012). Pour ce faire, nous avons identifié deux états, codés de 0 à 1: « 0 » absence de l'entrée ; « 1 » présence de l'entrée. 183 séquences distinctes ont été isolées. Parmi les 70% les plus fréquentes, quatre montrent l'introduction de mots nouveaux (38% du corpus) notamment en 1823, 1855, 1873 et 1908, tandis qu'une met en évidence le socle d'entrées présentes sur l'ensemble de la période (6%). Une autre témoigne de la continuité entre le Bégin et les Nysten-Littré (7%). Enfin, une dernière montre la filiation entre les Nysten-Littré (31%). Ces différents éléments ont été synthétisés ci-après sous forme tabulaire:

Dictionnaires	**En effectifs**	**En %**
Lavoisien (1793)	4.606	4
Capuron (1806)	6.414	6
Bégin et al. (1823)	11.215	10
Nysten (1833)	9.308	8
Littré/Robin (1855)	16.755	15
Littré/Robin (1865)	18.940	17
Littré/Robin (1873)	22.366	20
Littré-Gilbert (1908)	22.583	20
Total des entrées	**112.187**	**100**

Table 1: Description générale du corpus
Indication de lectur: La nomenclature du Lavoisien contient 4.606 entrées soit 4% du corpus

Entrées/dictionnaires	**En effectifs**	**En %**
Un	14.348	38
Deux	4.825	13
Trois	5.653	15
Quatre	4.158	11
Cinq	1.854	5
Six	2.736	7
Sept	1.416	4
Huit	2.375	6
Ensemble	**37.365**	**100**

Table 2: Répartition des entrées suivant leur présence dans les dictionnaires
Indication de lecture: Sur un total de 37.365 entrées dédoublonnées, 2.375 sont présentes dans l'ensemble des dictionnaires, soit 6% du corpus

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Les deux dictionnaires les plus anciens représentent 10% du corpus, soit l'équivalent du Bégin. Au sein des différentes éditions du *Dictionnaire de Nysten*, entre 1833 et 1855, près de 80% des entrées ont été ajoutées ou renouvelées ; 13% en 1865 ; 18% en 1873 avant de tomber à 1% en 1908. Ceci traduit une progressive stabilisation de la nomenclature médicale.

Afin de confirmer ces résultats, nous avons également réalisé une analyse des correspondances multiples (ACM).[18] Il s'agit d'une technique descriptive et exploratoire visant à résumer l'information contenue dans un nombre quelconque de variables, ici six actives et deux supplémentaires, afin de faciliter l'interprétation des liaisons qui peuvent exister entre elles (Baccini 2010, p. 27). Notre ACM en tableau disjonctif complet[19] est construite d'après les résumés numériques des variations des nomenclatures des six dictionnaires de la lignée. Les coordonnées du Bégin et du Lavoisien sont projetées *a posteriori* à des fins comparatives. Afin de déterminer le nombre de facteurs à retenir pour la construction du plan factoriel, Raymond B. Cattell propose d'étudier la courbe de décroissance des valeurs propres soit le pourcentage d'inertie restitué par chaque axe. L'idée est de détecter les « coudes », les cassures qui marqueraient un changement de structure.[20] Cette approche est intéressante parce qu'elle permet de dépasser un arbitraire purement numérique.[21] Ici, la première dimension contient 44,6% de l'information, la deuxième 19,4%, la troisième 15,7%, la quatrième 9,3%, la cinquième 7,1% et la sixième 3,9%. Un décrochement est par ailleurs observé à partir de la deuxième dimension suivi d'un décroissement progressif. Nous avons donc retenu les deux premières dimensions qui représentent 64% de l'inertie cumulée soit de la dispersion des variables.

À présent, traçons l'ACM des variables suivant leur cosinus^2 (voir figure 1 ci-après). Cette mesure fournit la qualité de la représentation des modalités sur chaque axe. Elle est généralement associée au pourcentage de contribution, c'est-à-dire à la part occupée par la modalité sur l'axe considéré. Plus la variable est prépondérante, plus ces deux indicateurs seront élevés. Ce type de représentation graphique fait ressortir aussi bien les contributions les plus élevées que les plus marginales. Ici, le premier axe est structuré par les modalités « ne pas appartenir à la nomenclature du Littré/Robin de 1865 » et « y figurer » de coordonnées -0,87 et 0,85. Leur contribution respective à l'axe est de 14,2% (cos^2 0,75) et de 13,8% (cos^2 0,75). Le deuxième axe est quant à lui structuré par les modalités « ne pas appartenir à la nomenclature du Capuron » et « y figurer » de coordonnées -0,34 et 1,63. Leur contribution respective à l'axe est de 8% (cos^2 0,55) et de 39% (cos^2 0,55).

[18] Pour une présentation plus détaillée, voir le support pédagogique d'Alain Baccini.

[19] Les données se présentent sous la forme d'un tableau de Burt soit une « juxtaposition de tables de contingence où seules les liaisons entre variables prises deux à deux sont considérées. Il s'agit en *statistique d'interactions d'ordre deux* ». (Baccini 2010, p. 28).

[20] D'un point de vue théorique, Cattell préconise de ne sélectionner que les facteurs qui précèdent le coude observé (Cattell 1966). Il révise ensuite sa position, et décide d'intégrer le facteur du coude (Cattel/Vogelmann 1977). En réalité, tout dépend de la valeur associée.

[21] Le critère de Kaiser-Guttman ne considère que les valeurs propres supérieures à 1. S'il est préconisé pour les ACP ou AFC qui admettent des variables quantitatives, il est peu adapté avec des variables qualitatives. Ici, les valeurs propres sont comprises entre 0,04 et 0,4.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

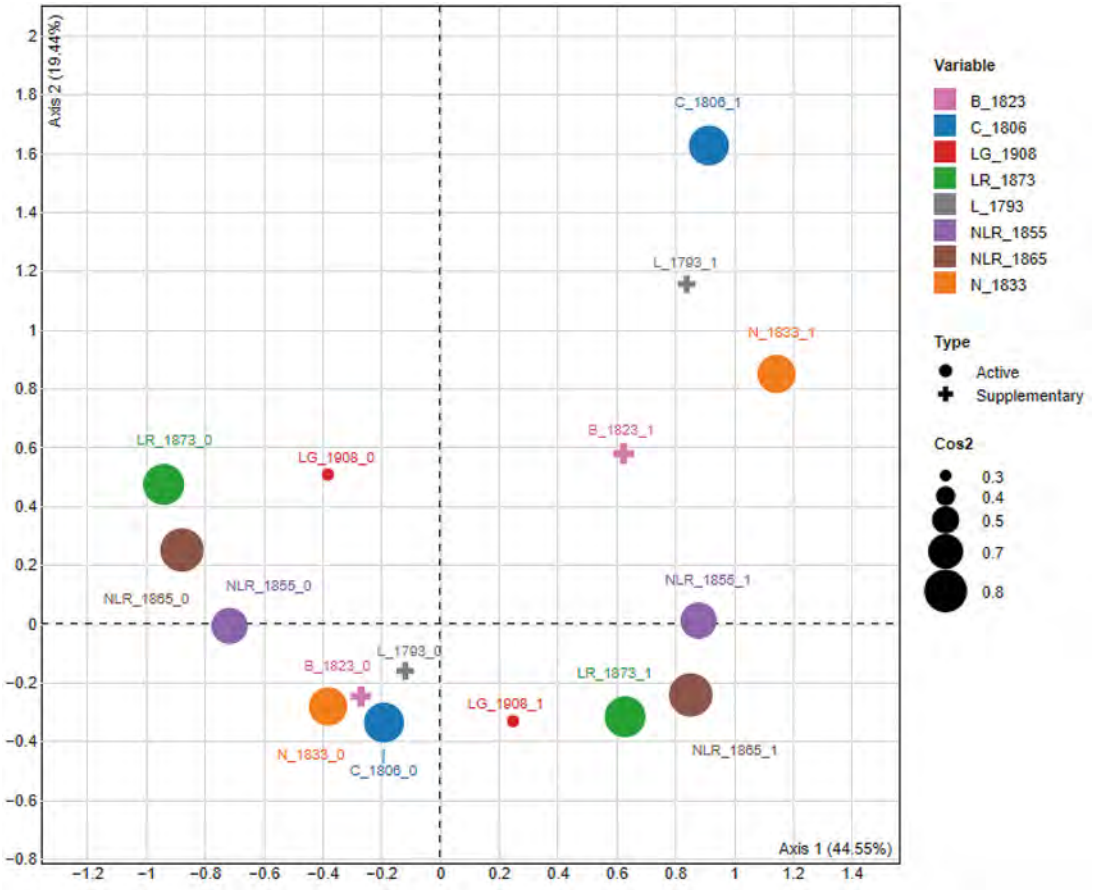


Fig. 1: ACM des variables suivant leur cosinus2

Si le graphique permet de discriminer l'absence/présence des termes dans la nomenclature des dictionnaires étudiés, la rupture incarnée par 1855 contribue à hauteur de 23,5% à la construction du premier axe, et divise le plan factoriel en deux. L'axe 1 regroupe ainsi les dictionnaires les plus récents tandis que l'axe 2 concentre les dictionnaires les plus anciens, plus particulièrement le Capuron (39%) et le Nysten de 1833 (15,5%). Les « positivistes scientifiques » sont donc opposés aux « spiritualistes ». On relève par ailleurs une proximité factorielle réciproque entre la nomenclature du Littré/Robin de 1865 (28%) et celle de 1873 (22%). Nous pouvons également admettre que les termes absents de la nomenclature du Lavoisien, du Capuron, du Bégin et du Nysten de 1833 sont similaires.

6. Conclusion

« Avec l'accroissement des faits, l'accroissement des termes ; avec la révolution des choses, la révolution des mots » (Dechambre 1864, p. XXXIV). Cette aventure lexicographique de la médecine française prend fin à l'aube du XXe siècle. Par son étendue et les polémiques qu'il a suscitées, le *Dictionnaire de médecine* y tient une place considérable autant par les nombreux étudiants qui l'utilisèrent que par la discipline du positivisme scientifique qu'il a su répandre dans une médecine en pleine expansion (Sournia 1981, p. 234). Sous l'impulsion de Littré, il prend l'ampleur d'un « véritable monument scientifique. C'est *le* Dictionnaire ! » (Littré-Gilbert 1908, p. VI). Si l'absence d'un terme est significative, la présence reste à voir. Il est effectivement nécessaire de dater le changement de sens s'il y a lieu. Pour confirmer

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

les premiers résultats de cette étude préliminaire, il sera donc nécessaire de l'élargir aux autres termes doctrinaux que nous avons pu relever.

Références

(1833): Bibliographie. Dictionnaire de médecine, de chirurgie, de pharmacie, des sciences accessoires et de l'art vétérinaire, de Nysten, 1833. In: Gazette médicale de Paris: journal de médecine et des sciences accessoires, 2 (1–19), p. 124.

(1872): Revue hebdomadaire. Académie des sciences: le Dictionnaire de médecine de MM. Littré et Robin, 1872. In: Gazette médicale de Paris: journal de médecine et des sciences accessoires 51 (4), p. 617.

Baccini, A. (2010): Statistique descriptive multidimensionnelle (débutants). Toulouse. http://www.math.univ-toulouse.fr/~baccini/zpedago/asdm.pdf (dernier accès : 04/05/2022).

Bégin, L.-J. et al. (1823): Dictionnaire des termes de médecine, chirurgie, art vétérinaire, pharmacie, histoire naturelle, botanique, physique, chimie. Paris.

Briand, J./Bricheteau, I./Henry, E. O. (1833): Dictionnaire de médecine, de chirurgie, de pharmacie, des sciences accessoires et de l'art vétérinaire de P.-H. Nysten. 5e édition. Paris.

Bricheteau, I. (1830): Statistique médicale. In: Moreau, J.-L. (dir.): Encyclopédie méthodique, médecine, 13 SEM-Z. Paris, p. 113.

Capuron, J. (1806): Nouveau dictionnaire de médecine, de chirurgie, de physique, de chimie et d'histoire naturelle, où l'on trouve l'étymologie et l'explication des termes des sciences, avec deux vocabulaires, l'un grec, l'autre latin, et les Synonymies relatives aux anciennes et nouvelles nomenclatures. Paris.

Cattell, R. B. (1966): The scree test for the numbers of factors. In: Multivariate Behavioral Research 1 (2), pp. 245–276.

Cattell, R. B./Vogelmann, S. (1977): A comprehensive trial of the scree and KG criteria for determining the number of factors. In: Multivariate Behavioral Research 12 (3), pp. 289–325.

Chaumeton, F. P. (1814): Dictionnaire. In: Dictionnaire des sciences médicales par une société de médecins et de chirurgiens, 9 DES-DIS. Paris, pp. 268–282.

Dagognet, F. (1997): Positivisme. In: Ambrière, M. (dir.): Dictionnaire du XIXe siècle européen. Paris, pp. 937–938.

Dechambre, A. (1869) : Lavoisien. In: Dictionnaire encyclopédique des sciences médicales 2 (2), LAR-LOC. Paris, p. 106.

Dechambre, A. (1879): Nysten. In: Dictionnaire encyclopédique des sciences médicales 2 (13), NEZ-NYS. Paris, pp. 853–854.

Desrosières, A. (1993): La politique des grands nombres. Paris.

Duryer, A. (1824): Réflexions sur les dictionnaires de médecine. Paris.

Giraud, L. (1862): L'histoire d'un livre: le Dictionnaire de médecine de P.-H. Nysten, revu par Émile Littré et Charles Robin. Paris.

Gouhier H. (1999): Bergson et le Christ des Évangiles. Paris.

Guardia, J. M. (1865): Bibliographie. Dictionnaire de médecine, de chirurgie, de pharmacie, des sciences accessoires et de l'art vétérinaire, d'après le plan suivi par Nysten, Littré/Robin 1865. In: Gazette médicale de Pari : journal de médecine et des sciences accessoires 3 (20), pp. 203–206; 218–220.

Lavoisien, J.-F. (1793): Dictionnaire portatif de médecine, d'anatomie, de chirurgie, de pharmacie, de chymie, d'histoire naturelle, de botanique et de physique qui contient les termes de chaque art, leur

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

étymologie, leur définition et leur explication, avec un vocabulaire latin et françois, et un grec-latin et françois, à l'usage de ceux qui lisent les auteurs anciens. Nouvelle édition corrigée et augmentée. Paris.

Littré, E./Robin, C. (1855): Dictionnaire de médecine, de chirurgie, de pharmacie, des sciences accessoires, et de l'art vétérinaire, de P.-H. Nysten. 10e édition. Paris.

Littré, E./Robin, C. (1865): Dictionnaire de médecine, de chirurgie, de pharmacie, des sciences accessoires et de l'art vétérinaire d'après le plan suivi par Nysten. 12e édition. Paris.

Littré, E./Robin, C. (1873): Dictionnaire de médecine, de chirurgie, de pharmacie, de l'art vétérinaire et des sciences qui s'y rapportent. 13e édition. Paris.

Littré, E./Gilbert, A. (1908): Dictionnaire de médecine, de chirurgie, de pharmacie et des sciences qui s'y rapportent. 21e édition. Paris.

Maupertuis, P.-L. (1793): Statistique. In: Hassenfratz, J.-H.: Encyclopédie méthodique, physique 4. Paris, p. 612.

Moreau, J. (1976): L'homme et son âme, selon Saint Thomas d'Aquin. In: Revue philosophique de Louvain 74 (21), pp. 5–29.

Robette, N. (2012): L'analyse de séquences: une introduction avec le logiciel R. https://quanti.hypotheses.org/686 (dernier accès: 21-03-2022).

Savary, A. C. (1810): Nouvelles littéraires. In: Journal de médecine, de chirurgie et de pharmacie 20, pp. 393–399.

Sournia, J.-C. (1981): Littré, lexicographe médical. In: Revue de la Société française d'histoire de la médecine 15 (3), pp. 227–234.

Informations de Contact

Anaïs Chambat
Université Paris Cité
Direction générale déléguée des bibliothèques et musées (DGDBM)
Bibliothèque interuniversitaire de santé – pôle médecine
ana.chambat@gmail.com

Remerciements

Cette recherche a été effectuée dans le cadre du projet CollEx-Persée « Métadictionnaire médical multilingue » de la bibliothèque numérique française Medica porté par la Bibliothèque interuniversitaire de santé – pôle médecine, l'Université Paris Cité, Sorbonne Université, l'Institut universitaire de France et l'Université de Lorraine. Je tiens à remercier mes collègues Nathalie Rousseau et Jean-François Vincent pour leurs relectures attentives.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Mihai-Alex Moruz/Mădălina Ungureanu

17TH-CENTURY ROMANIAN LEXICAL RESOURCES AND THEIR INFLUENCE ON ROMANIAN WRITTEN TRADITION

Abstract This paper focusses on the first Slavonic-Romanian lexicons, compiled in the second half of the 17th century and their use(rs), proposing a method of investigating the manner in which lexical information available in the above corpus relates, if at all, to the vocabulary of texts from the same period. We chose to investigate their relation to an anonymous Old Testament translation made from Church Slavonic, also from the second half of the 17th century, which was supposed to be produced in the same geographical area, in the same Church Slavonic school or even by the same author as the lexicons. After applying a lemmatizer on both the Biblical text (Books of Genesis and Daniel) and the Romanian material from the lexicons, we analyse the results and double the statistical analysis with a series of case studies, focusing on some common lexemes that might be an indicator of the relatedness of the texts. Even if the analysis points out that the lexicons might not have been compiled as a tool for the translation of religious texts, it proves to be a useful method that reveals interesting data and provides the basis for more extensive approaches.

Keywords 17th century; Church Slavonic; bilingual dictionaries; Old Testament; Old Romanian

1. Introduction

1.1 An overview on the beginnings of the Romanian lexicography

In Romanian culture, as was generally the case, the starting point of lexicographical practice can be identified in the vernacular glosses on texts written in the culture language of the time, according to the following stages: glosses → glossaries → bilingual dictionaries (Adamska-Sałaciak 2014, p. 1; see also Kovalenko 2016, p. 275, for the Russian culture context; Gruszczyński/Saloni 2013 for Poland). Drafting bilingual dictionaries is a direct consequence of the contact between two languages: in the case of the Romanian first dictionaries (17th century) the language of culture (Latin in Transylvania, Church Slavonic in Moldavia and Wallachia), used also in administration, blended with Romanian and the vernacular language tended to assimilate the status of the other language. The Romanian glosses on Slavonic texts illustrate the first stage of this contact. The brief rudimentary glossaries following the model of similar Slavonic works mark the second stage: two such works dating from the 16th century are known (see Strungaru 1966, p. 146; Mihăilă 1972, p. 308; Gînsac/Ungureanu 2018, p. 847, n. 4). In both cases, the words are not listed in alphabetical order. Large bilingual dictionaries, with words arranged alphabetically, appear in the third stage (17th century). This category includes bilingual lexicons, such as Anonymus Caransebesiensis, a Romanian-Latin dictionary dating from the mid-17th century; a Latin-Romanian one, from the end of the same century; an Italian-Romanian lexicon compiled a few years before 1700 and a trilingual dictionary, Latin-Romanian-Hungarian, compiled at the end of the 17th century as well. This category also includes the six Slavonic-Romanian dictionaries which are the object of our research.

We can thus note that most of the dictionaries issued at the time are of the L2-L1 type (in which L2 is either Latin, Italian, Hungarian or Slavonic) and only one of these dictionaries is L1-L2 (Romanian-Latin); referring to the purpose of the latter, Chivu (2008, p. 34) states

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

that it "aimed at the most detailed presentation of the Romanian language with the help of Latin glosses", as opposed to L2-L1 dictionaries that seem to have been working tools used in translation (Chivu 2012, p. 45), both of liturgical texts and of administrative documents (Gherman 2021, p. 2). We shall focus our analysis on the Slavonic-Romanian dictionaries dating from the 17th century.

1.2 The features of the Slavonic-Romanian lexicons from the 17th century

All six Slavonic-Romanian dictionaries were compiled in the second half of the century (except for Lex.Mard., dating from 1649), in the same geographical area (Wallachia), following the model of the Slavonic-Ruthenian lexicon published in 1627 in Kyiv by Pamvo Berynda, a reference work for the 17th century, containing around 7,000 entries grouped in two alphabetical lists, i. e. a list of old common names and a list of proper names and terminology borrowed from Hebrew, Greek and Latin (Stankiewicz 1984, p. 152). Popular at the time, this lexicon played an essential role in the Ukrainian, Romanian, Russian, Byelorussian, and Polish lexicography (ibid., p. 52). Along with the Slavonic Grammar written by Meletie Smotrițki (1619), this dictionary was used as an instrument for learning Slavonic in Petru Movilă's Academy in Kyiv (Ševčenko 1984, p. 22), and thus it was only natural to serve as a model for Slavonic-Romanian dictionaries. All six dictionaries are preserved as manuscripts, two of them in Russian libraries and the remaining four at the Romanian Academy Library in Bucharest (see their complete list in the Bibliography section). They are works of large dimensions, preserved almost completely (the leaves corresponding to letter A are missing from Lex.3473, and a few leaves from the same letter are also missing from Lex.Staicu). These works are obviously related (with the possible exception of Lex.Mard.), most likely being processed copies of a unique Slavonic-Romanian intermediate work that was either lost or not yet found (see Gînsac/Ungureanu 2018, p. 872).

Of these works, only Lex.Mard. has been edited so far, by Crețu (1900). The other lexicons have been studied only partially and solely with regard to a series of specific aspects (the connections between the six lexicons established by analysing certain fragments; hypotheses regarding their paternity; their connection with Lex.Ber.; see Strungaru 1966; Mihăilă 1972; Gînsac/Ungureanu 2018, p. 850). The lack of interest in editing these lexicons derives most probably from the idea that they are pretty similar and tributary to their source (see, for instance, Chivu 2012, p. 45). Ever since the end of the 19th century, several researchers have suggested the idea to elaborate a comparative edition (Mihăilă 1972, p. 324), in order to better illustrate the relations between these lexicons, on the one hand, and their connections with the Ruthenian model, on the other.

On a closer look, however, one may observe that these lexicons do not follow the model as faithfully as previously thought (see, for instance, Gînsac/Ungureanu 2018; Gînsac/Moruz/Ungureanu 2021), and each of them actually includes an innovative component. As a general observation, some of the entries from Lex.Ber. are omitted and new entries are added, but these new entries are not always the same in all six Romanian lexicons. Furthermore, the content of the definitions is modified. We must also mention that, despite being a bilingual dictionary, Lex.Ber. does not only render the Ruthenian equivalents of the Slavonic terms, but also indicates polysemous terms, figurative meanings, usage contexts (by quotations), bibliographic references, explanations. Romanian lexicons have a general tendency of simplifying the definitions in the model, yet in some cases extensive explanations are provided.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

One of the challenges in studying the Slavonic-Romanian lexicons is that they have not been edited so far; the parallel study of six large texts, with a rich lexical inventory, is a difficult task. Starting from this idea, a comparative digital edition has been created in which the entries are aligned at headword level, allowing parallel viewing (see http://www.scriptadacoromanica.ro/bin/view/eRomLex/ and also Gînsac/Moruz/Ungureanu 2021). Besides being the first lexicographic works in the Romanian culture and thus an important part of the Romanian cultural heritage, the 17th century Slavonic-Romanian lexicons contain a vast amount of lexical material that has to be inventoried and exploited in the Romanian language thesaurus (DLR), in works on the history of the Romanian language, etc. This corpus is also a valuable resource for the study of the lexicographical activity of the time from the standpoint of its social and cultural function: its analysis in relation to other writings or translations of the time can yield clues with regards to the authorship of certain texts.

2. The relation between these lexicons and other writings of the time

2.1 Why were the lexicons compiled?

In this study we shall regard the dictionaries as "tools in an integrated network" (Varantola 2002, p. 30), focusing on a specific element of this network, namely the context in which these dictionaries were produced (by "context" we understand the other texts produced in the same period, within the same geographical area), for the study of which we propose a specific work method. The study of this context may undoubtedly lead to the identification of the type of users the dictionaries targeted and also to the identification of their authors. Given the lack of paratextual elements,[1] the identification of the other elements of the functional network that the 17th-century Slavonic-Romanian dictionaries belong to[2] is based on the analysis of the lexicon contents and the cultural context in which they were produced.

As far as the purpose of the first bilingual Romanian dictionaries is concerned, we should note that they could have been connected to the activity of the Church Slavonic schools existing in Wallachia (where they could have served as didactic instruments) and also to the translation activity of both religious and administrative texts.[3] This idea was motivated by the (so far partial) analysis of the entry inventory. However, we believe that these hypotheses need to be validated by several methods: an exhaustive analysis of the entry inventory of the Romanian lexicons in relation to the inventory of the model (in other words: which entries from Lex.Ber. are preserved, which are ignored?); a qualitative analysis of the defini-

[1] The title is actually mentioned in just two dictionaries: Lex.Mard. and Lex.Pet.; the author's name is recorded only in Lex.Mard., whereas Lex.3473 has a name mention that could belong to the copyist and in the case of Lex.Staicu the author is only presumed based on the orthography and the very brief study of the vocabulary, see Strungaru (1966).

[2] See Lew (2015, p. 2): "Dictionary use occurs in a particular context, and users reaching for dictionaries are typically immersed in a particular kind of activity".

[3] In turn, Lex.Ber. would have been compiled as a tool for a new edition of the Slavonic Bible (as the existing edition, the Ostrog Bible, 1580, was perceived as being too obscure), on the initiative of Feodor Balaban, Berynda's patron. In any case, Berynda did not limit his approach to the Biblical text, but consulted a much more diverse corpus (see V. Nimciuk, Introduction to Lex.Ber., online: http://litopys.org.ua/berlex/be01.htm, accessed on 09-03-2022); further on, V. Nimciuk argues that Lex.Ber. was used not only for didactic purposes but also by a wide range of readers of Slavonic texts, for whom the Lexis published by Lavrentii Zizanyi in 1596 was no longer adequate.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

tions (how do the authors of the lexicons process the information in the source?); a statistical analysis of the Romanian definitions' content (what words did they use?), by comparison with the same type of analysis applied to other texts from the same period. In our study we shall focus on the last of these strategies, analysing the relation between the lexical content of these dictionaries and the content of a significant work in the Romanian culture of the time, also translated into Romanian from Slavonic, namely a translation of the Old Testament, from the same period and the same geographical area as the lexicons.

2.2 Premises

The first texts preserved in the Romanian language date from the 16th century and are mainly versions of the Psalters and Apostles and other religious texts, most of them translated from Slavonic and some from Hungarian. During the 17th century, the source languages also include Greek and Latin. Religious translations are still most numerous (liturgical, homiletics, canon law texts and catechisms, etc.), mainly from Slavonic, yet other works are also issued, such as civil law works, popular books and even some original texts (chronicles and poetry). The first translation of the New Testament is published in 1648, followed in 1688 by the first full translation of the Bible in Romanian (henceforth: B 1688). As for the Old Testament, the first translations into Romanian date from the second half of the 17th century: a translation from Greek (henceforth: OT 45), and a second translation having as sources the Church Slavonic Bible from Ostrog, a Latin version published in Anvers and the abovementioned translation from Greek (henceforth OT 4389)[4].

The presumed author of OT 4389 is Daniil Panoneanul (Andriescu 1998, p. 14), a professor of Church Slavonic (before 1660) at the school in Târgoviște, in Wallachia (Ursu 1995, 2003), translator and subsequently Bishop. Furthermore, based on certain linguistic particularities, Ursu (2003, p. 198) advances the hypothesis that Daniil Panoneanul could have been the professor of Staicu Grămăticul, presumed author of one of the six Slavonic-Romanian lexicons (Lex.Staicu) and that Daniil Panoneanul could actually have been the author of a lexicon that was in fact lost and from which all the other preserved lexicons derived, except for Lex.Mard. (the hypothesis of a Romanian lexicon which could have been the source for the other five was also formulated in Gînsac/Ungureanu 2018). The hypothesis regarding the common authorship of the translation of the Old Testament from Church Slavonic and the lost Slavonic-Romanian lexicon should be verified by comparing the lexical inventory of this OT version and the Romanian lexicons; it is the aim of this article, it being a first step (and a proof of concept) for a broader investigation into the relatedness of these writings and also into the purpose for which these lexicons were compiled (as instruments for translation from Church Slavonic, to be used as a teaching tool in Slavonic language schools, or simply in the act of reading Slavonic texts). Our hypothesis is that a common authorship could be proven by common lexemes, especially if these were specific to a small geographic area, and by common translation choices.

[4] This translation is kept in Rom. ms. no 4389 from the Library of the Romanian Academy in Bucharest. All three translations were edited within the "Monumenta linguae Dacoromanorum. Biblia 1688" series (Iași, "Alexandru Ioan Cuza" University Press, 24 vol., 1988-2015). For further information on this version, see Ursu 1995; Ursu 2003; Cândea 1979, pp. 129–134. The OT 4389 translation is dated 1665–1672 (Cândea 1979, p. 131); for comparison, Lex.Staicu was dated 1669 (Strungaru 1966) or cca 1660–1670 (Mihăilă 1972, p. 313); Lex.Pet.: around 1693 (Creţu 1900, p. 50) or ante 1693 (Mihăilă 1972, p. 315); Lex.3473: 1672-1673 (according to two marginal notes); Lex.1348: 1678 (Mihăilă 1972, p. 313).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

2.3 Analysis

2.3.1 Methodology

In order to extract the relevant static data for our intended analysis, both OT 4389 and the lexicons themselves had to be pre-processed. Firstly, OT 4389 was available in rich text format, in an unstructured file. The text needed to be separated into separate biblical books, and then transformed into the appropriate XML format for the lemmatizer and POS tagger. In our initial runs, we used the current structured form of the lexicons (i. e., without lemmatization of the glosses), but in later iterations we have also added the sequence of lemmatized words from the gloss.

For each of the books, we have then searched for each autosemantic word in the glosses from each of the lexicons. Initially, we intended to forego the lemmatization of the lexicons, seeing that, in most cases, they do not contain full sentences but rather disparate words and phrases, and, as such, a POS tagger might not yield accurate results. In order to account for variations on word forms, we have employed a set of similarity measures (such as Levenshtein or Cosine), but on analysing the results, we determined that the matches obtained were poor. We had, therefore, to also lemmatize the Romanian glosses in the lexicons, regardless of the low accuracy of obtained data, and, upon checking the results, we have found that this choice yielded significantly better accuracy. Below, we have given our analysis on two of the biblical books, Genesis and Daniel.

We chose to perform the comparison on the Book of Genesis (= Gen.) because the text is longer, the lexis is quite diverse and also because it is mentioned several times as a reference in the body of definitions in Lex.Ber. (and also in the Romanian lexicons). Another characteristic is the high number of proper names, yet this particular difficulty was overcome by eliminating them manually. We opted for the elimination of proper names so that we could avoid providing an erroneous perspective on the relation between the lexis units that made the object of our analysis. The fact that the authors of Romanian lexicons retain just a small number of the proper names included in Lex.Ber. is already known. Moreover, the authors of Romanian lexicons do not retain ethnonyms, the names of Biblical peoples. The few proper names and ethnonyms that still occur are included in the definitions of other words.

The same operations were performed for the Book of Daniel (= Dan.), which we chose because it is from the second part of OT 4389, it is shorter, and it is a prophetic book, the discourse having other linguistic characteristics than in Genesis. We aimed at verifying the validity of the results obtained for Gen. on another book.

The outcomes of automatic lemmatization required manual correction. For instance, in one gloss from Lex.Pet., the lemma "a îngreca" ('to get pregnant') is extracted for the form "îngreez", while the actual lemma is "a îngreuna" ('to make difficult'). This is a case in which automatic processing results in 'false friends' that need to be eliminated manually for the sake of statistical accuracy.

Initially, prior to the elimination of proper names and ethnonyms, 1348 autosemantic words were identified in Gen. OT 4389 following automatic lemmatization. Subsequently, after the inventory analysis and the elimination of proper names and ethnonyms, there remained 1135 autosemantic words. Of these, 542 are found in Lex.Mosc., 773 in Lex.Staicu, 859 in Lex.3473, 779 in Lex.Pet., 558 in Lex.1348, and 718 in Lex.Mard. A number of 167 words (around 10%) from Gen. OT 4389 were not found in any of the lexicons. The percentage

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

would have been much higher (around 30%), had we not eliminated the above-mentioned categories from the inventory.

Following automatic lemmatization, 770 autosemantic words were found in Dan. OT 4389. We eliminated the proper names, the Hebraisms (e.g., *mane*, *thekel*, *fares*), the ethnonyms. There remained 735 autosemantic words. Of these, 539 are to be found in Lex.Mosc.; 539 in Lex.Staicu; 597 in Lex.3473; 541 in Lex.Pet.; 418 in Lex.1348; 496 in Lex.Mard. 84 lexemes from Dan. OT 4389 were not found in any of the lexicons.

2.3.2 Interpretation of results

A first observation refers to the absence from the lexicons of the proper names occurring in Gen. OT 4389. We have selected those rare instances in which proper names were preserved, but, in the great majority of cases, Biblical proper names and names of Biblical peoples (occurring quite frequently in the Book of Genesis: Philistines, Canaanites, Egyptians, etc.) are omitted in the Romanian lexicons. For instance, *Eghipet* occurs in the lexicons in the explanation for the Slavonic лавириѳъ, therefore with no reference to the Biblical text (in fact, it was an extremely well-known proper name); however, *eghiptean* 'Egiptian' has no occurrence whatsoever in the lexicons. *Iacov* occurs in lexicons in the definition for бенонѣ (the name of one of Jacob's sons), which is a reference to Gen. 25, 18. This anthroponym was also very common. *Iordan* (Gen. 32, 10) also occurs in a definition (for акриды – the name of a plant); *Isac* – in the definition for the Slavonic глумлюсѧ 'to walk', as a reference to Gen. 24[5].

One can note that the lexical inventory common to Gen. OT 4389 differs from one lexicon to another. Not surprisingly, the highest degree of commonality seems to be a feature of Lex.3473, which can be explained by the fact that this lexicon provides more extensive Romanian definitions; things are quite different in the case of Lex.1348, which most often indicates only the Romanian equivalent of the Slavonic headword, without further details, even when the definition in Lex.Ber. is extensive. The high degree of commonality between Gen. OT 4389 and Lex.Mard. is quite surprising, as the latter has a smaller inventory of entries compared to the other lexicons (see Gînsac/Moruz/Ungureanu 2021, pp. 5 f.); however, the author of Lex.Mard. does not innovate with regard to the lexicon inventory, remaining the most faithful to the source (Lex.Ber.). Furthermore, we should note the small number of terms in Gen. OT 4389 which do not appear in lexicons; this aspect may indicate either that the texts originate from the same dialectal area or that their authors shared the same basic vocabulary. The 168 terms missing from the lexicons can be explained by the very nature of the texts, as these terms are specific to Biblical texts (e.g., *tîrîtoare* 'crawlers', *slujnic* 'servant'); on the other hand, the lemmatizer separated inflectional forms (mostly participles) that have no equivalent in the lexicons because of the specificity of their (descriptive) definitions (for instance, *a spăla* 'to wash' occurs in the lexicons, whereas *spălat* 'washed' does not). Other words (e.g., *cinie* 'tool', *a se ciudi* 'to wonder') could have already had an archaic character in the second half of the 17th century; their presence in the translation of the OT

[5] Whenever the context is indicated in the lexicons, it can be compared with its translation in Gen. OT 4389. In this particular case, OT 4389: "Şi ieşi Isaac îndeseară la cîmp să se primble" (= Isaac went out early in the evening to the field to walk). Lexicons: "Ieşi Isaac a se primbla la cîmp îndesară" (= Isaac went out to the field to meditate early in the evening). The differences are related to word-order (the translation from OT 4389 does not follow the word order in the Slavonic text, as opposed to the lexicons, in which word order is faithfully preserved).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

is due to the conservative nature of the religious language, while the lexicons were not governed by the same constraints.

The above observations are also valid for the Book of Daniel: the greatest number of common lexemes is to be found in Lex.3473, and the smallest – in Lex.1348, with a surprisingly high percentage in Lex.Mard. (given its relatively small number of entries). Some of the 84 lexemes missing from the lexicons are specific to the Biblical texts (*căpetenie* 'chief, ruler', *cioplitură* 'pagan idol', *greșeală* 'sin', *osîndă* 'punishment', a *pîngări* 'to defile', *prorocie* 'prophecy', etc.). Others might have already been considered archaic in the second half of the 17th century (*a conceni* 'to destroy', *concenire* 'destruction'), specific to the religious texts, which were based on a tradition, but not to a dictionary that used the current language for equivalences and definitions. Equally interesting is the case of *a blagoslovi* 'to bless' (from. Slavon. благословити), found in OT 4389, but missing from the lexicons, which give, for this exact Slavonic word, the equivalent *a binecuvînta*, calquing the Slavonic model, but with components of Latin origin.

We have given below a few lexemes that can be found both in OT 4389 and in the lexicons. We have correlated this information, whenever possible, with the occurrences in the Romanian language thesaurus (DLR) and compared it with the other version of the OT mentioned above (OT 45, which originated from a different geographic area, namely Moldavia) and with the 1688 Bible.

The noun *filosof* 'philosopher' is found in the OT 4389 in Gen. 41, where it is the equivalent of the Slavonic сказателъ. This Slavonic term is equated in the Romanian lexicons by *spuitor* ('the one who speaks'), an etymological rendering (сказати = *a spune* 'to speak'). However, the term *filosof* existed in the vocabulary of those who compiled the first lexicons, since it is used in a definition (for бгъ 'God'), where it translates Slavon. любомѫдрьцъ (literal: 'wisdom lover'). For this word DLR does not provide an exhaustive distribution of occurrences, but indicates as the first attestation a work dating from 1642. In any case, this was not a frequently used word.

The noun *posadnică* 'mistress' is common for the lexicons and the OT 4389; in all the cases, it equates the same Slavonic term, заложница. In the other two texts (OT 45 and B 1688), in the same context it is used the word *țiitoarea* (also to be found in the OT 4389). In the DLR, the term is considered regional, the first attestations dating from the same period in Wallachia.

An equally interesting term is *venetic* 'foreigner' (occurring both in the lexicons and the OT 4389), for which the OT 45 uses *nemearnic*. According to the DLR, this term is also attested in Wallachia in the second half of the century.

The noun *pușcărie* 'jail' is used both in lexicons and in OT 4389; in lexicons it is used as the equivalent of Slavon. темница – this choice is interesting, since Rom. *temniță* was also available. According to DLR, the first attestation of *pușcărie* is in the second half of the 17th century, in texts from Wallachia, while *temniță* had been used since the 16th century. Also intriguing is the choice of the translator of OT 4389, who uses *temniță* and *pușcărie* in the same verse (Gen. 39, 22), both as translations of Slav. темница – probably in order to avoid repetition, while in the next verse *pușcărie* is used twice, wich means that he used *temniță* due to the Slavonic word he had to translate, but *pușcărie* was more familiar to him.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

3. Conclusions

The first conclusion derived from the comparison of the lexical inventory from OT 4389 (the Books of Genesis and Daniel) with the one in the lexicons is that these lexicons do not seem to have been conceived as translation tools or at least as translation tools for the Biblical text. The lack of specific terms, the absence of proper names and of names of Biblical peoples, the selection of the terms that seems to favour the fundamental vocabulary, all these aspects point to a type of user that was not necessary a clergyman; to this we may add the observation that the definitions that are rather explicative indicate the fact that the lexicons were more likely drafted for text reception rather than text production.

We noted a few cases that could indicate a relation between the texts we analysed; however, we have insufficient proof to claim a common paternity of these texts; for relevant results, the investigation must be extended towards the entire Biblical text. Another possible approach would be the reverse one, in which the inventory of the lexicons would be related to the inventory of the OT. Last but not least, quantitative analysis is necessary, as it facilitates data extraction; nevertheless, it has to be correlated with a qualitative analysis that would focus on the Slavonic terms equated in each case (thus a translational approach to the resulting material must be added). The statistics, however, indicate interesting data that will provide the basis for a broader future approach.

References

Corpus

Lex.1348 = The Lexicon from Rom. ms. no. 1348, Romanian Academy Library, Bucharest (1–84v).

Lex.3473 = The Lexicon from Rom. ms. no. 3473, Romanian Academy Library, Bucharest (1–369).

Lex.Ber. = Pamvo Berynda: Леѯиконъ славеноросскїй и именъ тлъкованїе, Kiev, 1627. Edition by V. Nimciuk, 1961. Online: http://litopys.org.ua/berlex/be.htm.

Lex.Mard. = The Lexicon of Mardarie Cozianul, Rom. ms. no. 450, Romanian Academy Library, Bucharest, 1649.

Lex.Mosc. = The Lexicon from Moscow, Russian State Archive of Old Documents, Fond 188, Оп. 1. Ч. 2., p. 491.

Lex.Pet. = The Lexicon from Petersburg, the National Library of Russia in Sankt Petersburg (notice n° Q.XVI.5 – Славяно-молдавскийсловарь).

Lex.St. = The Lexicon from Rom. ms. no. 312, Romanian Academy Library, Bucharest (41r–216v).

OT 4389 = The Old Testament from Rom. ms. no. 4389, Romanian Academy Library.

OT 45 = The Old Testament from Rom. ms. no. 45, Library of the Romanian Academy, Cluj-Napoca.

Secondary literature

Adamska-Sałaciak, A. (2014): Bilingual lexicography: translation dictionaries. In: Hanks, P./de Schryver, G. M. (eds.): International handbook of modern lexis and lexicography. Berlin/Heidelberg, pp. 1–11. https://doi.org/10.1007/978-3-642-45369-4_6-1.

Andriescu, Al. (1988): Locul Bibliei de la București în istoria culturii, literaturii și limbii române literare. In: Andriescu, Al. et al.: Monumenta linguae Dacoromanorum. Biblia 1688. Pars I. Genesis. Iași, pp. 7–45.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Cândea, V. (1979): Raţiunea dominantă. Cluj-Napoca.

Chivu, G. (2008): Dictionarium valachico-latinum. Primul dicționar al limbii române. Bucharest.

Chivu, G. (2012): Lexiconul de la Buda, primul dicționar modern al limbii române. In: Analele Universității "Alexandru Ioan Cuza" din Iași. Secțiunea IIIe. Lingvistică 58, pp. 45–56.

Creţu, G. (ed.) (1900): Mardarie Cozianul. Lexicon slavo-românesc şi tîlcuirea numelor din 1649. Bucharest.

Gherman, A. M. (2021): Lexicography and the history of culture (The case of Teodor Corbea's Romanian-Latin dictionary). In: Diacronia 14, A189 (1–11), https://doi.org/10.17684/i14A189en.

Gînsac, A.-M./Ungureanu, M. (2018): La lexicographie slavonne-roumaine au XVIIe siècle. In: Zeitschrift für romanische Philologie 134 (3), pp. 845–876.

Gînsac, A. M./Moruz, M. A./Ungureanu, M. (2021): Slavonic–Romanian lexicons of the 17th century and their comparative digital edition (the eRomLex project). In: Diacronia 14, A192, pp. 1–11, https://doi.org/10.17684/i14A192en.

Gruszczyński, W./Saloni, Z. (2013): From multilingualto monolingual dictionaries. A historical overview of Polish lexicography. In: Studies in Polish Linguistics 8 (4), pp. 205–227.

Kovalenko, K. (2016): Sixteenth- and seventeenth-century Russian lexicons: pecularities of the first representatives. In: Studia Slavica Hungarica 61 (2), pp. 275–283.

Lew, R. (2015): Dictionaries and their users. In: Hanks, P./de Schryver, G. M. (eds.): International handbook of modern lexis and lexicography. Berlin/Heidelberg, pp. 1–9. https://doi.org/10.1007/978-3-642-45369-4_11-2.

Mihăilă, G. (1972): Contribuţii la istoria culturii şi literaturii române vechi. Bucharest.

Romanian Academy (2010): Dicționarul limbii române. Bucharest (= DLR).

Ševčenko, I. (1984): The many worlds of Peter Mohyla. In: Harvard Ukrainian Studies 8 (1–2), pp. 9–44.

Stankiewicz, E. (1984): Grammars and dictionaries of the Slavic languages from the middle ages up to 1850: an annotated bibliography. Berlin.

Strungaru, D. (1966): Începuturile lexicografiei române. In: Romanoslavica 13, pp. 141–158.

Ursu, N. A. (1995): Activitatea literară necunoscută a lui Daniil Andrean Panoneanul, traducătorul Îndereptării legii (Târgoviște, 1652) (I). In: Studii şi cercetări lingvistice 46 (1–6), pp. 157–173.

Ursu, N. A. (2003): Activitatea literară necunoscută a lui Daniil Andrean Panoneanul, traducătorul Îndereptării legii (Târgoviște, 1652) (IV). In: Studii şi cercetări lingvistice 54 (1–2), pp. 189–201.

Varantola, K. (2002): Use and usability of dictionaries: common sense and context sensibility?. In: Corréad, M. H. (ed.): Lexicography and natural language processing. A festschrift in honour of B. T. S. Atkins. Grenoble, pp. 30–44.

Contact information

Mihai-Alex Moruz
Faculty of Computer Science, "Alexandru Ioan Cuza" University, Iaşi
mmoruz@info.uaic.ro

Mădălina Ungureanu
Institute of Interdisciplinary Research, Department of Social Sciences and Humanities, "Alexandru Ioan Cuza" University, Iași
madandronic@gmail.com

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Acknowledgements

This work was supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI – UEFISCDI, project number PN-III-P1-1.1-TE -2019-0517, within PNCDI III.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Clarissa Stincone

USAGE LABELS IN BASNAGE'S *DICTIONNAIRE UNIVERSEL* (1701)

Abstract Basnage's revision (1701) of Furetière's *Dictionnaire universel* is profoundly different from Furetière's work in several regards. One of the most noticeable features of the dictionary lies in his increased use of usage labels. Although Furetière already made use of usage labels (see Rey 1990), Basnage gives them a prominent role. As he states in the preface to his edition, a dictionary that aspires to the title of "universal" should teach how to speak *in a polite way* ("poliment"), *right* ("juste") and making use of specific terminology for each art. He specifies, lemma by lemma, the diaphasic dimension by indicating the word's register and context of use, the diastratic one by noting the differences in the use of the language within the social strata, the diachronic evolution by indicating both archaisms and neologisms, the diamesic aspect by highlighting the gaps between oral and written language, the diatopic one by specifying either foreign borrowings or regionalisms.

After extracting the entries containing formulas such as "ce mot est …", "ce terme est …" and similar ones, we compare the number of entries and the type of information provided by the two lexicographers[1]. In this paper, we will focus on Basnage's innovative contribution. Furthermore, we will try to identify the lexicographer's sources, i. e. we will try to establish on which grammars, collections of linguistic remarks or contemporary dictionaries Basnage relies his judgements.

Keywords Historical lexicography; *Dictionnaire universel*; Basnage de Beauval; 17th century; usage labels

1. The dual role of quotations and usage labels in Basnage's *Dictionnaire universel*

Henri Basnage de Beauval is the reviser of Antoine Furetière's monumental *Dictionnaire universel* (1690), the second monolingual dictionary published in France. Basnage's *Dictionnaire universel* (1701) is profoundly different from Furetière's work in several regards. Two essential elements in the revision lie in his increasing quotations and usage labels. Quotations are marginal in Furetière's dictionary and "y répond avant tout à des intentions littéraires" (Lehmann 1995, p. 49). On the contrary, in Basnage's dictionary they play a prominent role, being in all the entries for which the lexicographer was able to find them.[2] Usage labels, on the other hand, are used by Furetière more extensively than quotations (see Rey 1990). However, they appear to be extremely few when compared to the amount of usage labels recorded by Basnage. Both elements have the role of illustrating the *bel usage* of words. This had already been the main purpose of the *Dictionnaire de l'Académie française.* Its lexicographers, although they chose not to quote, had taken the effort to mark those words "qui commencent à vieillir, & ceux qui ne sont pas du bel usage, & que I'on a qualifiez de bas ou de style familier" (Preface to the *Dictionnaire de l'Académie*, 1694). On his part, Furetière had set out to create a universal dictionary that would contain terms specific to

[1] The .txt files digitised with Transkribus and subsequently analysed with BBEdit are flawed (incorrectly separated words, confused or missing letters, etc.). It is possible that some usage labels may have escaped analysis.

[2] Most terminological entries lack quotations because terms are rarely used by writers.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

arts and crafts without giving much weight to the *bel usage*[3]. Basnage tries to bring together the two lexicographical visions. While maintaining Furetière's universalist spirit,[4] Basnage depicts the *bel usage* of the French language by means of quotations and usage labels. However, "les nombreuses marques d'usage fournissent une multitude d'informations qui dépassent une conception uniquement puriste" (Gemmingen-Obstfelder 1982, p. 131). As shown elsewhere (see Stincone 2021), Basnage's lexicographic work could be regarded also as a proto-learner's dictionary since it guides its reader in understanding all aspects of the language. Indeed, while illustrating the *bel usage* of words, Basnage takes the reader by the hand, suggesting which words he can use without hesitation and at the same time warning him about the words to keep from being used. He accomplishes both purposes by means of a well-organised architecture of quotations and usage labels.

In principle, quotations and usage labels are mutually exclusive: if a word is in general use and it is in the texts of good contemporary authors, its status does not generate perplexity and quotations guarantee its *bel usage* while providing illustrative examples of the headword as well as syntactic models to reproduce for the reader. The presence of one or more usage labels, on the other hand, indicates that the word cannot be found in the texts of good contemporary authors, that not all language experts approve of the word or that for some reason it is excluded from the *bel usage*. The reader should therefore be careful not to use it. Some entries contain both, usage labels and quotations. This is the case with the entries referring to the burlesque, comic and satirical literature. Furthermore, if Basnage is aware of a single good author who used the word, his quotations can be preceded by formulas such as "[Auteur] (l') a employé", "[Auteur] s'(en) est servi", "Ce mot se trouve dans [Auteur]". This last one, in which the author's name designates his work by metonymy, is also used to indicate the lexicographer[5] who records the lemma when it is not present in the works of contemporary writers. These formulas accomplish the primary goal of ensuring the existence of the word. Finally, when Basnage is dubious about the *bel usage* of a word for which he cannot find attestations in the works of contemporary authors or in his reference dictionaries, he inserts a usage label, introduced by "On doute que", which explains his doubts.

2. Quotations

Quotations, almost exclusively from works of the second half of the 17th century, accompany the dictionary's definitions. They are extracted from various types of works, primarily literary texts but also historical and religious ones, correspondence, scientific and informative treatises, informative journals.

Basnage is not the first French lexicographer who used quotations. Before him, Pomey had used them in his bilingual French-Latin dictionary (see Girardin 1995) and Richelet (see Lehmann 1995) in what is to all intents and purposes the first monolingual French dictionary. Basnage takes quotations from both. Olivier Patru, jurist and writer but also academi-

3 The very rare occurrences of "bel usage", "beau langage" and "beau stile" in Furetière's *Dictionnaire universel* are mostly found in definitions or illustrative examples rather than in usage labels while Basnage uses them in twenty-seven, eighteen and six usage labels respectively.

4 In addition to integrating the descriptions of almost all the dictionary entries, Basnage introduces new ones belonging both to the common lexicon and to specialised languages.

5 Basnage, in his revision, systematically consults the monolingual French dictionaries of Richelet (16932 [1680]) and Académie française (1694) as well as the bilingual French-Latin dictionaries of Nicot (1606), Pomey (1671), Danet (1673) and Tachard (1689).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

cian, clarifies Richelet's working method in a letter to Maucroix where he asks him to strip his texts and those of Balzac while "Richelet va dépouiller tout d'Ablancourt" (Livet 1858, pp. 50 f.) and Patru himself will look for quotations in his pleas. Patru reports Richelet's idea that at least five or six living authors for the sake of being quoted themselves, would also provide quotations from others. Richelet's dictionary is therefore a collective work containing quotations identified by several collaborators. In addition to the forty-nine authors mentioned in the alphabetical table at the beginning of the dictionary, "il faut y ajouter les auteurs vivants tels Patru, La Fontaine, Benserade, Dépréaux et d'autres encore" (Lehmann 1995, p. 41). All these authors' quotations in Basnage's dictionary are very often a clue allowing easy access to the source from which Basnage transcribes his examples, i. e. Richelet's dictionary. Most likely, Basnage, in addition to taking up Richelet's quotations, also takes up his working method and gets help from some collaborators in stripping the texts of contemporary authors in order to identify the appropriate quotations for each entry. Basnage's quotations are very numerous and extremely varied. They will not be treated in this article.

3. Usage labels

Many different types of usage labels in Basnage's *Dictionnaire universel*, although appearing distant and unrelated to each other, are part of a vast and coherent project.

The usage labels allow Basnage to point out which senses deviate from what he considers to be the correct way of speaking and writing, relying on the opinions of language experts and grammarians who had expressed their views on words' usage. In the 17th century there is the proliferation of the "genre grammatical des réflexions et remarques sur la langue" (Leroy-Turcan, 1998, p. 90). In particular, Vaugelas's *Remarques sur la langue française* (1647) inspired the works of Boisregard, Bouhours, Corneille, Ménage and Tallement, which are incorporated in the second edition of the *Dictionnaire universel*. These lexicographers, grammarians or language enthusiasts are so reputable that they can *approuver* or *desapprouver*, *admettre* or *condamner* the use of a word in the same way as Richelet and the lexicographers of the *Académie française*. Some are purists, others are liberal with regard to words and constructions to be admitted or banned. Since the French language is in the process of being defined, the approval of one language expert does not imply that of another. In general, Basnage records all the positions of which he is aware, expressing his preference from time to time. The above-mentioned verbs are found in 78 usage labels of Basnage's edition, which often devotes entire paragraphs of the entry to linguistic disquisitions.

The usage label is called NOTTE by lexicographers of the classical age. As we read in the two editions of the *Dictionnaire universel*, "dans un Dictionnaire on doit mettre une *notte* à un mot, quand il est vieux, ou particulier à quelque art ou science. Quand il est dans l'usage commun, il n'y faut point de *notte*." (NOTTE, DU1690 and DU1701).[6] This entry shows that the usage labels adopted by both Furetière and Basnage are of two types: the usage labels that, placed after the headwords, detect the specialised languages by indicating the domain to which the entry belongs, and the usage labels that, placed at the end of the entry, refer to the dimensions of linguistic variation. Basnage significantly increases the ones and the others. Although indicated by means of various formulations, the domain is mostly introduced by "(en) terme(s) de [name of domain]". On the one hand, Basnage introduces a large num-

[6] We reproduce Basnage's spelling for sentences appearing in both editions.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

ber of headwords and sub-headwords, marking them as specialised lexicon if it is the case while, on the other hand, he adds the domain to which already existing entries belong. In this article we will not consider domain designations[7] and we will focus on linguistic variation. Although the entry NOTTE mentions only archaisms, Basnage's *Dictionnaire universel* describes a much wider spectrum of the linguistic variation. The lexicographer does not neglect the description of any dimension of language, namely diachronic, diatopic, diastratic, diaphasic, and diamesic. Unfortunately, the description of the different dimensions is not systematic and uniform. An attempt has been made to identify recurrent patterns in order to extrapolate significant data. Usage labels are introduced by "ce mot" in a certain number of entries, by the personal pronoun "il" in others, by impersonal periphrases such as "on dit" or "on s'en sert" in still others. Sometimes, they refer to all senses of the word, sometimes to the last sense only. In some cases, Basnage specifies it by means of the complement of limitation "en ce sens", in some he does not.

The discourse on the usage labels is extremely complex and cannot be covered in one article. For instance, we are not going to deal with usage labels referring to pronunciation, syntactic constructions, words' figurative and connotative meaning, to the lexicographer's opinions on frequency of use, to experts' doubts about words' usage, to popular common mistakes, to the type of texts and contexts adopting certain words, to words' collocations and so on[8]. Furthermore, the various dimensions are often intertwined. For the sake of discussion, we have chosen to analyse separately the five axes of linguistic variation in Basnage's *Dictionnaire universel*, whose data are always compared with those of Furetière's edition.

3.1 The diachronic dimension

The diachronic dimension refers to the linguistic variation linked to time.

One of Furetière's illustrative examples of VIEUX ("Ce mot est *vieux*, il n'est plus du beau langage", VIEUX, DU1690), is the prototype of a very frequent usage label in the dictionary. It shows that the "age" of the word is a parameter which establishes its exclusion from the "beau langage", elsewhere designated as "beau stile" or "bel usage". In the 16th century, Malherbe, in his quest for linguistic purity, launched a war against archaisms that was not yet over by the 17th (see Brunot 1909). In spite of the opposition of authoritative linguists to the censorship of archaisms, purity is a shared aspiration. As an example, Bouhours "compared 'beautiful language' to pure, clear water that had no taste, in other words, it contained no archaisms" (Cormier 2008, p. 164).

Basnage considerably increases the number of usage labels referring to archaisms[9] in his edition compared to the previous one. Furetière's usage labels marking archaisms are 150 while Basnage's 314. Specifically, Furetière states that a word "est vieux" in 86 entries and that "vieillit" in 62. He uses the periphrasis "commencent à vieillir" only once referring to a specific sense of a couple of headwords and his only use of the past participle "vieilli" concerns non-lemmatised words. The range of formulations used by Basnage is wider. The lexicographer introduces a kind of degree of word's "ageing". He marks 137 archaisms by

7 Domain designations of Basnage's *Dictionnaire universel* are partially analysed in Galleron/Williams (in print).

8 Some of these issues are partially treated in Stincone (2021).

9 The number of archaisms does not coincide with the number of usage labels marking archaisms. This article does not take into account the numerous definitions opening with "Vieux mot".

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

means of the adjective "vieux" in predicative function after the verb *être*, for instance "Ce mot est vieux" (SEMONCE, DU1701). The verb *vieillir*, conjugated exclusively in the present indicative by Furetière, is flexed 177 times by Basnage, both in the past and in the present indicative, within 21 and 133 usage labels respectively which indicate that the word has already been banned from use or that the banning is not yet concluded, e. g. "Ce mot a vieilli" (PARFAIRE, DU1701) and "Il vieillit, & ne se dit guere qu'en riant" (LARMOYANT, DU1701). Plus, a periphrastic construction used in 23 entries by Basnage projects the decay into the future, like in "Le mot d'*indice* en ce sens commence à vieillir" (INDICE, DU1701). Sometimes, Basnage marks the intensity of the verb by means of adverbs or adverbial locutions such as *absolument*, *fort*, *un peu* and *trop* which further define the expressive climax. All of these usage labels are inspired by those contained in Richelet's and *Académie*'s dictionaries as well as by the remarks of language experts.

Basnage does not just report archaisms. In an effort to provide a tool that is as inclusive as possible, he welcomes newly formed words into his dictionary. Furetière marks only three words as neologisms.[10] One of them gives the date of introduction of the word into the French language: it is 1684 (see SURTOUT, DU1690 and DU1701), six years before the publication of the *Dictionnaire universel* which Furetière had finished in 1688. It is, therefore, a word of only four years old when Furetière compiles his dictionary. Basnage does not suppress the indication continuing to consider "nouveau" a word coined fifteen years before and marks other nineteen words as neologisms. Some of them are not well integrated into the language system since the lexicographer states that they are not yet established, e. g., "Ce mot est nouveau, & n'est pas encore tout-à-fait établi" (RHETORICATION, DU1701). Sometimes, a neologism can only be used in low or familiar styles, sometimes it is even "trop nouveau pour s'en servir" (INVESTISSEMENT, DU1701). The usage labels of a few entries, inspired by Danet's dictionary which is explicitly mentioned, concern the lemmatization of the newly conceived feminine equivalents of masculine nouns, formed by means of a Latin-derived suffix (CONDUCTRICE, CONTEMPLATRICE, CORRECTRICE). Maybe in the circles of the *Precieuses*, derided by Molière in the *Precieuses ridicules* (1659), the all-modern need was felt to endow the language with feminine designations.

3.2 The diatopic dimension

The diatopic dimension refers to the linguistic variation linked to the geographical location of the speakers. It is described at a broad level already by Furetière, who is aware that "par toutes les Provinces le peuple parle un *jargon* different de la langue des honnêtes gens" (JARGON, DU1690 and DU1701). In this sentence two linguistic axes overlap, the diatopic one and the diastratic one being the language of "provincial people" opposed to that of "honest people". In fact, the language described by both lexicographers is that of the more literary and educated milieu of the court and those who have relations with it, not that of common people[11]. The Paris region, essentially because of the presence of the *Académie française* and of Versaille, was the area to which most of the literati converged. It is as if the

[10] This does not mean that Furetière does not record neologisms. According to Rey (2006, pp. 162 f.), Furetière's *Dictionnaire universel* contains almost all the 17th century neologisms listed by Brunot (1909) in his list of "mot nouveaux".

[11] Basnage absorbs and makes his own the thought of Vaugelas whose definition of "bon usage" is: "C'est la façon de parler de la plus saine partie de la Cour, conformement à la façon d'écrire de la plus saine partie des Autheurs du temps" (Vaugelas 1647, p. 18).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

court at Versailles exudes an air of politeness around it, so that "quand ce provincial aura *humé* l'air de Paris, il sera plus poli" (HUMER, DU1701). The provincial, i. e. the one who comes from the province or who lives in the province, "c'est un homme qui n'a pas l'air, & les manieres de la Cour ; qui n'est pas poli ; qui ne sçait pas vivre ; qui n'a point vu le monde" (PROVINCIAL, DU1701). Basnage recognises the existence in France of "12. principales Provinces" (GOUVERNEMENT, DU1701) but he does not give a list of them[12]. Nevertheless, plenty of entries contain references to the language spoken in one or more of the French provinces such as Anjou, Bourgogne, Bretagne, Champagne, Normandie, Picardie, Provence, etc. which are not equally represented in the dictionary.

Specifically, the contrast between Paris and the province emerges in eight entries, two of which are already in Furetière. Three of them state that the word is in use in the province rather than in Paris while the remaining five provide two different variants, one for Paris, one for the province. More often, Basnage refers generically to the language of the province without mentioning Paris. The reference to the language of the "province(s)" is contained in 85 sentences, 38 of which are already used by Furetière. These occurrences are of two types: 16 of them record the province variant as opposed to the word used in Paris while the remaining 69 contain a usage label which restricts the word's usage to the province(s), for instance: "On ne se sert de ces mots-là qu'en certaines Provinces" (ADMODIER, DU1701).

Sentences referring to borrowings will not be analysed here as the etymological notations do not direct the reader on the word's usage.

3.3 The diastratic dimension

The diastratic dimension refers to the extraction and social status of the speakers.

As previously seen, Basnage's *Dictionnaire universel* essentially describes the language of the "honest people". In general, the reference to the language of the people is sufficient to mark a deviation from the *bel usage*. Basnage emphasises that a word is used by the "honnêtes gens" only to point out some contrasts with the language of the "peuple", e. g., "Ce mot ne se dit ordinairement que par le peuple de Paris; car les honnêtes gens disent un Vendeur de melons" (MELONNIER, DU1701). The diastratic dimension is indicated in the dictionary in many ways. The most recurrent adjectives referring to people's language are "bas" and "populaire" which characterise the style "dont use le peuple" (STILE, DU1701).

Of the 411 entries in which one or more usage labels containing references to the "bas" style appear, only 70 are already in Furetière's dictionary. If we exclude the 82 entries associating the adjective with burlesque, comic and satirical literature and the 57 entries associating it to the familiar register, which will be discussed later, 272 entries still remain. Sometimes, Basnage inserts a usage label which relegate all senses or, more often, only the last one to the low style immediately after the illustrative examples provided by Furetière, e. g. "Il a mangé tout son crevé saoul. Cette derniere expression est basse." (CREVER, DU1701).[13] In these cases, the usage label is based on the lexicographer's perception of the language. In other occasions, Basnage introduces a new sense specifying that it belongs to the low style: "On dit, Ruer en *cuisine*, pour dire Goinfrer. Il est bas." (CUISINE, DU1701). In these cases,

[12] Both lexicographers mention "Normandie, […] Bretagne, […] Anjou" (PROVINCE, DU1690 and DU1701).

[13] The underlined text is already in Furetière's edition.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

the sense as well as its usage label is generally taken from Richelet's or the Academy's dictionaries.

Moreover, 235 entries contain references to the "populaire" (often associated to "bas") character of a word. Furetière's dictionary already records 150 of them. Generally, Furetière opens his definitions by specifying that the word is a "terme populaire" while Basnage's usage labels are mostly placed at the end of the entry and refer to the last sense described by Furetière, e. g., "Il lui a tant corné aux oreilles cette maxime, qu'enfin il l'en a persuadé. Il est bas & populaire" (CORNER, DU1701). Furthermore, in 64 entries, 47 of which are already in Furetière's *Dictionnaire universel*, an entire paragraph is introduced by formulas such as "On dit populairement" aimed at describing uses of the word interdicted to the *bel usage*. It should be noted that Richelet never uses the adjective *populaire* or the adverb *populairement* in his dictionary in order to describe a linguistic usage. Only eight times these usage labels are taken from the *Dictionnaire de l'Académie* while, generally, Basnage comments on Furetière's examples relying on his own linguistic awareness.

The language of the "(petit) peuple" is also described in 179 entries, 111 of which are already in Furetière's dictionary. As was the case in the previous section, they are not all usage labels. On one side, there are sentences containing the popular variant of the word, which is often a scientific term, while on the other side there are full-fledged usage labels containing a usage restriction to the "peuple" or even the "petit peuple" (in 39 entries). In particular, there are 67 entries recording popular variants, most of which (49) are already in Furetière's while a usage information is contained in 112 entries, over half of which (60) are in Furetière's. As an example of usage label, "Ce mot [...] n'est en usage que parmi le peuple" (BAGARRE, DU1701). These usage labels are often taken from Richelet's, more rarely from the Academy's dictionary. In some cases, they are Basnage's formulations.

3.4 The diaphasic dimension

The diaphasic dimension refers to the linguistic variation associated with the change in style and expressive register. The expressive register varies depending on the communicative situation. One can say something "par ironie", "par plaisanterie", "par mepris", all uses recorded by Basnage. Also, the adoption of a specific literary genre implies diaphasic variation. Basnage does not seem to perceive the difference between the diastratic dimension and the diaphasic one since he considers the low or popular style to be "celui dont use le peuple, ou dont on use dans le comique, ou le burlesque" (STILE, DU1701), thus equating lower social classes language with the language of writers who imitate common people's language. It seems appropriate to distinguish here the "low" language of illiterate people who have no alternative in expression from the one of authors who consciously choose to adopt a certain literary genre. In his edition, Basnage greatly increases references to burlesque, comic, and satirical literature.

Although they designate different literary genres, the adjectives *burlesque*, *comique*, and *satirique* are often used together and in conjunction with *bas* which confirms their belonging to the same style. The frequent reference to the *figuré* style is not surprising if one considers that the search for ambiguity and double meaning is the base of burlesque, comic and satirical production. A word does not belong to the *bel usage* to Basnage's eyes if it is exclusively found in works of one of the three literary genres. As previously said, entries containing a reference to one of them present a peculiar feature: they contain both usage labels and author quotations. Since "seuls les genres burlesques, comiques et satiriques peuvent

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

s'accomoder du mauvais usage" (Brunot 1909, p. 50), the adoption of one of them implies the use of the entries discarded by the purists.

Furetière uses the adjective *burlesque* and the adverb *burlesquement* in 85 entries, the adverb *comiquement* in only one, and never refers to satirical literature. Basnage, for his part, refers to burlesque literature in 322 entries, to comic literature in 149, and to satirical literature in 37. Generally, Basnage's numerous usage labels pertaining to the three genres are placed after the definition and any illustrative examples, but before quotations – which illustrate the genre – and the etymological note. As an example, "Brouet, se dit aussi d'un mechant potage: mais dans le stile comique & burlesque. Le Galant, pour toute besogne, avoit un *brouet* clair. La Font." (BROUET, DU1701). The few quotations by Furetière come from Colletet, Corneille, author of *L'Illusion comique*, Desmarests de Saint-Sorlin, author of the *Visionnaires,* Mairet, author of the *Sylvie*, Marot, author of the *Epigrams*, Molière, Pasquier, Regnier, Scarron, author of the *Virgile travesti*. With the exception of Desmarests, whose citation is suppressed, Basnage retains Furetière's quotations, adding others. Molière (most often "Mol.") is the most frequently cited author, followed by La Fontaine (most often "La Font." or "La Fon."), Scarron (most often "Scar."), Voiture (most often "Voit."), Saint-Amant (most often "St. Am." exceptionally "Saint Amand"), Boileau (most often "Boil." or "Boi."), Sarazin (most often "Sar."), Ablancourt (most often "Ablan."), Mainard (most often "Mai.", exceptionally "Main."), Regnier (most often "Reg."), Gombaut (most often "Gon." or "Gomb."), Benserade (also "Bens.")[14]. Basnage does not bother to read Molière or Scarron in order to find their quotations, almost all of them are extracted from Richelet's dictionary[15]. While Richelet often indicates his sources in a very precise way, providing in addition to the name of the author also that of the work and the page, Basnage is much hastier and, at the same time, systematic providing only the abbreviation of the author's name. Even though the tendency to downgrade the burlesque[16] takes hold as early as the 17th century, French authors seems to use it without hesitation. If Sarazin and Scarron are mentioned among the representatives of the burlesque genre (see BURLESQUE, DU1690 and DU1701), Molière among those of the comic (see COMIQUE, DU1690 and DU1701) and Boileau among the authors of satires (see SATYRE, DU1690 and DU1701), perhaps one might be surprised by the numerous quotations of Ablancourt, La Fontaine, or Voiture, elsewhere welcomed by Basnage among the "bon auteurs" of the French language, whose quotations guarantee the *bel usage* of the words. Basnage, therefore, does not condemn those authors for having used a low style, he only recognises that they have consciously chosen to use it.

3.5 The diamesic dimension

Closely related to the diaphasic dimension is the diamesic one, which concerns linguistic variation related to the physical-environmental medium through which communication takes place. Basnage greatly increases the number of usage labels that restrict the adoption of the word to the oral rather than the written.

[14] Molière is mentioned 70 times, Benserade 6. Other authors, mentioned in an occasional way, are not reported here.

[15] Although the spelling of Richelet and that of Furetière do not always correspond, Basnage can easily identify and then transcribe the quotations because Richelet adopts a labelling system which makes use of diacritical signs (see Bray 1990).

[16] See Nédelec (2004).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Out of the 80 usage labels containing a reference to "conversation", only five are already used by Furetière. About twenty of them are inspired by Richelet's and Academy's dictionaries while the remaining ones, formulated by Basnage, are based on the work of Boiregard, Bouhours, Caillières, Corneille, Ménage, Tallement, Vaugelas, whose surnames are often abbreviated at the end of the usage label. For instance, "Ce mot ne doit guere sortir de la conversation. Corn." (ENTACHER, DU1701). The noun *conversation* is followed by the adjective *familiere* in no less than eleven usage labels. As stated in the two editions of the dictionary, "le stile familier, est celui dont on se sert en conversation" (STILE, DU1690 AND DU1701). Therefore, the two words are strictly connected. The occurrences of *familier* are very numerous. Out of the 182 usage labels containing the adjective *familier*, only 20 of them are already used by Furetière. The remaining 162 are partly inspired by Richelet's and Academy's dictionaries, by some collections of linguistic remarks or are formulated by Basnage himself. Finally, the four usage labels containing the adverb *familierement* are all introduced by Basnage.

4. Conclusions

Basnage strongly increases the number of usage labels in his *Dictionnaire universel* (1701) by taking into account all aspects of linguistic variation. Usage labels have a dual role in the dictionary. In addition to illustrating the *bel usage* of the French language, they also clarify the linguistic doubts of users who are guided in choosing a word rather than another depending on the dimension related to the linguistic act. Prescriptivism and descriptivism seem to merge in Basnage's dictionary. On the one hand, the lexicographer feels the need to provide the prescriptions of the purists as well as the linguistic norm dictated by the *Académie française*. On the other hand, he feels the need to provide a tool that describes the language as it is, in every possible communicative situation and context. The lexicographer is sometimes inspired by the usage labels of Richelet's and Academy's dictionaries, sometimes he relies on the collections of remarks of French language experts of his time. He often trusts his own linguistic awareness as well while creating his usage labels.

References

Académie française (1694): Le Dictionnaire de l'Académie françoise dedié au Roy. Paris.

Bray, L. (1990): Les marques d'usage dans le Dictionnaire françois (1680) de César Pierre Richelet. In: Lexique 9/ Les marques d'usage dans les dictionnaires (XVIIe, XVIIIe siècles). Lille, pp. 43–59.

Brunot, F. (1909): Histoire de la langue française, des origines à 1900 ; 3, 1–2. La formation de la langue classique (1600–1660). Tome 3 (1). Paris. https://gallica.bnf.fr/ark:/12148/bpt6k58392786 (last access: 18-03-2022).

Cormier, M. C. (2008): Usage labels in The Royal Dictionary (1699) by Abel Boyer. In: International Journal of Lexicography 21 (2), pp. 153–171. https://doi.org/10.1093/ijl/ecn013 (last access: 18-03-2022).

DU1690 = Furetière, A. (1690): Dictionnaire universel. La Haye/Rotterdam.

DU1701 = Basnage de Beauval, H. (1701): Dictionnaire universel. La Haye/Rotterdam.

Galleron, I./Williams G. (in print): Tenir la promesse du Dictionnaire universel: l'esprit encyclopédique d'Henri Basnage de Beauval.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Gemmingen-Obstfelder, B. (1982): La réception du bel usage dans la lexicographie du XVII siècle. Actes du Colloques International de Lexicographie dans la Herzog August Bibliotek Wolfenbüttel (9–11 octobre 1979). In: Wolfenbütteler Forschungen (18), pp. 121–36.

Girardin, C. (1995): Une doctrine jésuite de l'exemple. Le Dictionnaire Royal Augmenté de François-Antoine Pomey. In: Langue française 106 (1), pp. 21–34. https://doi.org/10.3406/lfr.1995.6441 (last access: 18-03-2022).

Lehmann, A. (1995): La citation d'auteurs dans les dictionnaires de la fin du XVIIe siècle (Richelet et Furetière). In: Langue Française (106), pp. 35–54. https://www.jstor.org/stable/41558721 (last access: 18-03-2022).

Leroy-Turcan, I. (1998): Les grammairiens du XVIIe siècle et la première édition du Dictionnaire de l'Académie françoise. In: Le Dictionnaire de l'Académie française et la lexicographie institutionnelle européenne (Actes du Colloque organisé pour le troisième centenaire du Dictionnaire de l'Académie française). Paris, pp. 89–109.

Livet, C. (1858): Histoire de l'Académie française par Pellisson et d'Olivet. Paris.

Nédelec, C. (2004): Le burlesque au Grand Siècle: une esthétique marginale ? In: Dix-septieme siecle 224 (3), pp. 429–443.
https://www.cairn.info/revue-dix-septieme-siecle-2004-3-page-429.htm (last access: 18-03-2022).

Rey, A. (1990): Le marques d'usage et leur mise en place dans les dictionnaires du XVIIe siècle: le cas Furetière. In: Lexique 9 / Les marques d'usage dans les dictionnaires (XVIIe, XVIIIe siècles), pp. 17–29.

Rey, A. (2006): Antoine Furetière: Un précurseur des Lumières sous Louis XIV. Paris.

Richelet, P. (1680): Dictionnaire françois. Geneva.

Stincone, C. (2021): Le Dictionnaire universel de Basnage est-il aussi un dictionnaire d'apprentissage? In: La linguistique 57 (1), pp. 55–72.
https://www.cairn.info/revue-la-linguistique-2021-1-page-55.htm (last access: 18-03-2022).

Vaugelas, C. F. (seigneur de) (1647): Remarques sur la langue françoise, utiles à ceux qui veulent bien parler et bien escrire. Paris. https://gallica.bnf.fr/ark:/12148/bpt6k8707897j (last access: 18-03-2022).

Contact information

Clarissa Stincone
Université Sorbonne Nouvelle – Paris 3
clarissastincone@live.it

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Marija Žarković

THE LEGAL LEXICON IN THE FIRST DICTIONARY OF THE SPANISH ROYAL ACADEMY (1726–1739)

The Concept of the Judge

Abstract This paper consists of a short analysis of the sources and the treatment of the legal lexicon in the first dictionary published by the Spanish Royal Academy (1726–1739), followed by a longer commentary on the representation and the treatment of the concept of judge, in which the reflection of the extra-linguistic factors in the definitions stands in focus. The results highlight the relevance of the legal context of that era for the treatment of the lexicon related to the legal domain, but they also demonstrate the pattern in which the lexicographic data displays peculiarities of legal matters.

Keywords History of lexicography; legal lexicon; Spanish lexicography; Spanish Royal Academy

1. Preliminaries

The research presented below is part of a doctoral thesis *The history of the legal vocabulary in the dictionaries of the Royal Spanish Academy* (*1726–2014*).[1] This thesis aims to explore and interpret the lexicographical treatment of the legal vocabulary and the vocabulary related to the legal domain[2] in the first monolingual general language dictionary (the *Diccionario de autoridades* 1726–1739) and three posterior editions (the *DRAE* 1884, one of the editions published under the 20th century, and the *DRAE* 2014) published by the Royal Spanish Academy, and to investigate the extent to which the lexicographical data reveals not only the semantic-paradigmatic properties of the lexemes, but also the existing state of the language and the historical changes in the Spanish legal vocabulary. The study adopts a diachronic and comparative approach, in order to provide an overview of the semantic development of this lexicon while examining its treatment and prevalence in Spanish academic lexicography from the 18th century to the present day.

The study of the fragment presented in this article belongs to the part of the research devoted to the legal lexicon and the lexicon related to the legal domain in the *Diccionario de Autoridades* (in continuation *Autoridades*),[3] the first dictionary published by the Spanish Royal Academy (1726–1739), in a crucial era for both the history of law and the history of lexicography. This part of the study represents a starting point of the thesis and aims to set the ground for future comparative analysis of three posterior versions of the academic dictionary, followed by a comparative diachronic study of the findings.

---

1 This thesis is supervised by the Dr. Gloria Clavería Nadal (Autonomous University of Barcelona) and the Dr. Andreas Deutsch (University of Heidelberg). I would also like to thank Dr. Aniceto Masferrer (University of Valencia) for his valuable comments. The research is made possible by the Autonomous University of Barcelona and a grant from the Serbian Government which are gratefully acknowledged.

2 By "legal lexicon" I refer to the distinct specialised and semi-specialised lexicon used in both oral and written legal discourse, particularly in courtroom proceedings. By "lexicon related to legal domain" I mean the lexical units that partly belong to the general language but that denotate concepts closely related to the legal sphere.

3 For a more detailed study on the *Autoridades*, cf. Blecua (2006), Freixas (2010), Lázaro Carreter (1972), Ruhstaller Kuhne (2002) among many others.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

2. The legal domain in the *Autoridades*

In the first half of the 18th century, the Spanish justice system began a profound transformation that would continue through the following centuries, and whose consequences are in evidence to this day. It is in the midst of this reformation process that the Spanish Royal Academy was founded (1713) with the purpose to fixate the words of the Castilian language at its greatest state, and that the elaboration and the publication of the first dictionary of the Spanish Academy (1726–1739) took place. The treatment of the lexicon related to the legal domain in the *Autoridades* can, therefore, only be understood by recognizing the extralinguistic factors related to the legal context surrounding the elaboration and publication of this dictionary: on the one side, the Bourbon reforms, and the consequent inclination towards an absolutism with characteristics similar to the French, and, on the other side, the rationalism and humanitarianism of advancing Enlightenment thought.[4] Accordingly, the lexicon related to the legal sphere that was integrated into this dictionary captures a very peculiar moment in Spanish law history.

The fact that a general dictionary implements a significant number of specialised legal lexical units as an integral part of its microstructure shouldn't come as a surprise, since the gradual migration of certain specialized and semi-specialised lexical units into the general language, and the consequent admission of these units in the general language dictionaries, is a common consequence of the incorporation of specialized knowledge into the common knowledge.[5] The first Spanish academics were aware of this phenomenon and stated already in the forward of their dictionary the intention to include and respectively label the lexicon related to the practice of the courts of justice (*Autoridades* 1726, para. 10), even though the principal idea was to elaborate a common language dictionary.

Considering the fact that legal language is the product not only of the people who speak and write it, but also of the jurisdiction and the professionals that use it (Tiersma 2008, p. 8), the semasiological characteristics of the legal lexical units lemmatized in dictionaries at different points in the past vary accordingly.[6] As Cabré et al. (2011, p. 116) explain, not even definitions of terminological units in dictionaries are exempt from the influence of ideological modulators, even when these treat —allegedly objectively— scientific or legal issues. In this sense, the *Autoridades* offers a series of terms whose definitions allow the user to comprehend the imprint of the conception of justice of the early 18th century. However, I do not aim to offer here a panorama of the characteristic traits of legal science present in the first academic dictionary. My goal is rather to illustrate the treatment of the aforementioned lexicon, and to highlight the reflection of the juridical issues of that time in the lexicographic data devoted to the concept of judge, while focusing on the presence of extralinguistic factors in two aspects of the microstructure: labels and definitions.[7]

[4] The 18th century, the century of The Enlightenment, commenced with a change in the reigning royal house of the Hispanic Monarchy. When the king Carlos II, the last Habsburg Hispanic monarch, died without successors (1700), the Spanish crown passed to Philip V (1700–1746), descendent of the French House of Bourbon and one of the grandsons of the French Roi Soleil, Louis XIV, triggering the War of the Succession, which subsequently resulted in the territorial integration and legal unification of the monarchy.

[5] On the specialised lexicon in the general dictionaries, cf. Pérez Pascual (2012). On the case of legal terms in dictionaries, cf. Nielsen (2015).

[6] On the legal language and the reality, *cf.* Olivecrona (1999) and Baldinger (1985).

[7] To collect the entries for the analysis, I performed a manual extraction of the entries from the digital version of the *Autoridades* using two different methods. Firstly, for the gathering of the entries dealing

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

2.1 Sources of the legal lexicon

The creators of the *Autoridades* aimed to produce a cult reference work that would exhibit and preserve the Castilian language at its finest state and that would be based on the most respected literary and non-literary works. According to Freixas, the Greco-Roman tradition of including *autoridades* ('authorities'), *i. e.*, the sources of lexical data quoted in the final part of most of the articles in this dictionary, was decided by looking up to the dictionary published by Accademia della Crusca (1691) and allowed the Spanish academics to illustrate what they considered the proper use of lexical units (Freixas 2003, p. 33).

Just as the rest of the lexicon that forms part of this dictionary, the lexical units related to the legal sphere were extracted and exemplified by relying on different texts that the first Spanish Academy members had at their disposal.[8] The number and the variety of the texts used for the extraction of the legal lexicon demonstrate the considerable amount of documentation that were used as sources. Nevertheless, this doesn't mean that each of the entries in the dictionary comes from a textual source and includes an exemplifying quote. In fact, the largest part of all the entries that form part of the corpus of this research – almost 20% – is not related to any source, attesting the willingness of the academics to incorporate vocabulary specific to different professions even when they couldn't find an appropriate source to rely on, as was previously confirmed by the findings of M. Freixas (2010, p. 339).

The entries that do include a source of information show the high level of heterogeneity that was already demonstrated by the investigations on other types of specialised lexicon in *Autoridades*.[9] However, it is worth mentioning a disparity in the preference among the sources of legal lexicon. The entries that form part of the corpus of this research quote a total of 196 different texts, while only 29 of these are quoted in more than 10 entries. By far the most frequently quoted document is a code of law promulgated by King Philipp II in the 16th century, *La Nueva Recopilación de las Leyes del Réino*, quoted in almost 17% of the entries and subentries of the corpus (*alcalde de sacas* (*s. v. alcalde*), *calificador del santo oficio* (*s. v. calificador*), *assistente*). Closest to that, in terms of number of citations, is a law code *Las siete partidas* compiled during the reign of Don Alfonso X the Wise (1252–1284), that appears in 3,5% of the entries and subentries[10] (*s. v. merino*, *prueba*, *quebrantamiento*).

with the specialised legal lexicon I relied on the comments and indications of diatechnical usage and extracted the specialised legal lexical units that were marked as such. Secondly, in order to detect the lexical units related to legal domain that were not marked by the academics, I performed a search of a total of 80 key legal concepts prominent in the 18th century in the digital version of the dictionary. These legal concepts were previously gathered from relevant legal texts dating from the 17th and 18th century Hispanic Monarchy. Thus, the key concept "jurisdicción" ('jurisdiction'), referred to the entry *abad bendito* (s. v. *abad*). After a detailed analysis of the gathered terms, it was possible to separate those that actually belonged to the legal field from those that were primarily linked to some other sphere of the general terminology.

8 Many studies have been conducted so far on the sources implemented in the *Autoridades*, such as Lázaro Carreter (1972); Desporte (1998–1999); Freixas (2010, 2003, p. 412) estimated that there is a total of 460 writers quoted. For the legal sources in *Autoridades* in particular, *cf.* Freixas (2006).

9 Cf. For exaimple Gutiérrez Rodilla (1994–1995) on the study of the medical lexicon in *Autoridades*.

10 For a detailed overview of the legal documents in *Autoridades*, cf. Freixas (2006).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

2.2 The subject field labels

Even though the specialised vocabulary was not initially planned as a part of the *Autoridades*, many authentic specialised lexical units found their way in. The data corpus of this study contains a total of 489 entries and subentries marked as authentic legal lexical units using various references that provide some limitation of employment,[11] such as in cases of the articles *absolver de la instancia*, *arraigar*, *enormíssima*, *estelionato*, *evicción*, *lucro cessante* (*s. v. lucro*), *obligación antidoral* (*s. v. obligación*), *peculado*, *reivindicación*, *reivindicar*, *término ultramarino* (*s. v. ultramarino*), *etc.* On the other hand, a significant part of the corpus is made out of the lexical units denotating different concepts related to the legal domain that were not marked as legal lexical units, such as *abogacía*, *fiscal*, *sentencia*, *etc.*

The creators of the *Autoridades* recognized the relevance of indicating the peculiarities of use and employed different strategies in form of comments and abbreviations in order to inform whether the employment of a lexical unit is related to any social stratum – *vulgar*, *rústica*, *culta*, *etc.* –, or if it is exclusive to some geographical area – *usado en Andalucía*, *es mui común en Astúrias*, *Galicia*, *y la costa de Cantábria*, *etc.* – or, if it is a term belonging to a domain of specialized knowledge – *en Medicina*, *térm. de Música*, *etc.* As far as the legal lexicon is concerned, the most striking features of lexicographical treatment are the various procedures used by the Academy to indicate the legal nature of a lexical unit. The variations in lexicographical indications and labels reveal the lack of a regular procedure. When it comes to the degree of specialisation of a term, we can distinguish two main methods of labelling entries and subentries treating specialised and semi-specialised juridical lexical units that were used in the legal discourse by the professionals involved in the work of the courts of justice in 18th century Hispanic monarchy:

1) Firstly, the data gathered for the purpose of this research shows that the academics point out the lexical units used in what they designate as *estilo forense* ('forensic style').[12] These are, for the most part, subentries that treat general language lexical units which would adopt a specialized meaning when used in a legal context. According to the extracted data, there are 387 of these lexical units. The labelling of these was performed mainly by using the following indications: "en lo forense" (*s. v. presentarse*, *rebelde*, *suplicar*, *ver*), "en el estilo forense" (*s. v. conato*, *dar la cáusa por conclusa* (*s. v. concluso*), *fallar*, *pieza de autos* (*s. v. pieza*) *etc.*), "en lo jurídico" (*s. v. caso*, *ingenuidad*, *conjunto*, *variante etc.*), "en el derecho" (*s. v. ingenuo*, *enemigo*, *preterición*, *causas mayores* (*s. v. mayor*), *etc.*).

2) Secondly, there are 102 specialised lexical units marked as legal terms. These are distinguished by employing the following marks: "term. forense" (*s. v. caso negado*, *divisorio*, *interusurio*, *escriturar*), "término forense" (*s. v. auto*, *abrogacion*, *capitulaciones*, *recisión*), "term. jurídico" (*s. v. peculado*, *indotación*).

The relation of a lexical unit to a specific legal area is sporadically marked by shortcuts and sense indicators in the definitions, such as:

(1) DECRETO. **En el Derecho Canónico** es la constitución, o establecimiento que el Sumo Pontifice ordena o forma […] (*s. v. decreto*);

[11] On the questions of labelling in the *Autoridades*, cf. Blanco Izquierdo/Clavería (2019, pp. 340–346).

[12] As Henriquez Salido explains, with the indication *forense* the academics refer to the lexical units used in the professional activity of lawyers and the justice courts (2010, p. 155). For a detailed overview of the employment of the technical label *forense* inside the definitions of *Autoridades*, *cf.* Henriquez Salido (2010, pp. 157–163).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

(2) PRETERICIÓN. **En el derecho Civil** se entiende la omissión del que teniendo hijos herederos forzosos, no hace mención de ellos en su testamento […] (*s. v. preterición*);

(3) CUERPO DE DELITO. **En la Jurisprudencia criminal** es la señal, o vestigio que queda de haverse cometido el delíto, que sirve de principio y fundamento para su averiguación y castígo […] (*s. v. cuerpo*).[13]

These examples show that the creators of the *Autoridades*, at least in a sporadic manner, marked not only the difference between the technical and the general lexicon, but also between the lexical units used in a certain context and the specialised vocabulary. Moreover, they tended to occasionally specify the legal terms even more precisely by stating the legal branch in question. This does not insinuate, however, that each of the legal concepts was marked as such, nor that the rules were consistently applied.

2.3 The treatment of different legal areas

The lexicographical data gathered for the purpose of this study reveals, on one side, a wide variety of legal vocabulary, and, on the other side, unequal representation of different areas of law. The previously described diversity of the sources of lexical data employed by the academics justifies this inequality. To achieve a more complete vision of the varied range of concepts, I separated the extracted lexical units into different groups based on the legal area they are related to. Due to space limitations, only four groups of entries will be commented here.

The monopolization of the executive, legislative and judicial powers by the monarch is one of crucial elements of the society of the Old Regime, and it is perhaps best portrayed by the entry of the noun *decreto* ('decree'), defined as any order or determination of the king, in the matters related to justice, grace, or government. The medieval conception of a king as, above all, judge and of justice as a domain of the activity of royal power can be observed in the subentries of the entry *imperio*: *mero império*, defined primarily as the absolute power over the vassals embodied in the prince, and *mero mixto império* defined as the jurisdiction delegated by the prince to the lord of vassals or to the magistrates allowing them to judge and punish crimes, by imposing the corresponding corporal punishment.[14] As per García-Gallo (1971), the jurisdictional power, in particular, is considered one of the most important manifestations of sovereign power. This supremacy of the king not only over the judges, but also over the church, can be perceived in a comment in the entry *tuitivo*, *va* ('protective') that relates this adjective to a power the king has to lower the penalty inflicted by the ecclesiastical judges.

The legal concepts explicitly related the set of legal norms, promulgated, or recognized by the Catholic Church,[15] i. e., the canon law (such as *decreto*, *matrimónio spiritual* (*s. v. matri-*

[13] Every example quoted in this article was extracted from the digital version of the *Autoridades*. The bolded text was formatted by the author.

[14] "Mero império. El absoluto poder que reside en el Príncipe sobre sus vassallos […]" (s. v. *imperio*); "Mero mixto império. La jurisdicción comunicada por el Príncipe al Señor de vassallos o a los Magistrados, para juzgar las cáusas y castigar los delítos, imponiéndoles la pena corporal correspondiente […]" (s. v. *imperio*). The English versions of ancient Spanish texts included in this article have been translated by the author.

[15] A detailed overview of the ideology and the treatment of the religious lexicon in *Autoridades* is given in an article by Rodríguez Barcia (2004).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

mónio), *irregularidad*, *canones*, *etc.*), and to the concepts corresponding to the norms of the legal system of the monarchy that regulates the social dimension of the religious factor, i. e., the ecclesiastical law[16] (*beneficio curado* (*s.v curado*, *da*), *divorciar*) together with the concepts related to the court of the Spanish Inquisition (*calificador del santo oficio* (*s. v. calificador*), *inquisidor general* (*s. v. inquisidor*), *inquisidor*), are treated in a total of 113 different entries and subentries. The fact that canon and ecclesiastical law used to be synonyms until the Protestant Reformation in the 16th century, and that it was not before the 19th century that the distinction between two different branches of law begun to be clearly outlined (Mantecón Sancho 2018, p. 11), explains the synonymous use of the terms *canónico* and *eclesiástico* and the treatment of these two branches of law as one and the same within the *Ius ecclesiasticum*.

The influence of the harshness of the Old Regime's justice is the maybe most palpable in definitions of the entries devoted to different judicial actions related to criminal proceedings. It is true that some of the lexical units related to this field were lemmatized and defined by relying on ancient legal texts whose origins can be tracked to the Visigoth era. Consequently, these entries describe concepts that were outdated in the 18th century, demonstrating the willingness of the academics to preserve this archaic vocabulary in their dictionary.[17] Such cases are the ancient methods of proving innocence described in the articles *caldaria*, *compurgación* and *purgación vulgar* (s. v. *purgación*), that rely on superstitious practices and ancient gothic customs. Nevertheless, the means of interrogation marked in the *Autoridades* as authentic judicial methods actively used in that time often rely on subjecting the accused to torture: *tormentar*, dar *tormento* (*s. v. dar*), *questión de tormento* (*s. v. question*), and *tormento*. The phenomenon of social segregation in criminal processes is reflected in the entry of, for example, the noun *pruebas*, that is defined as a legal means of gathering evidence used particularly for proving noble lineage. On the other hand, it is in the 18th century when, influenced by the Enlightenment movement, the relevance of methods of proofs such as that of legal medicine significantly increased.[18] Even though the term "medicina legal" ('legal medicine') is not lemmatized in this very form, the *Autoridades* does capture this current in the entry *cuerpo de delito* (*s. v. cuerpo*) that provides an encyclopaedic definition of *corpus delicti* by approaching it from the point of view of criminal jurisprudence, describing in detail the practice of proving that a crime actually has been committed while founding examples on a criminal practice text from the late 17th century.

The cruelty of the punishments was not far behind. If we categorise the penalties in the corpus according to the legal good affected, we can see that there are twelve articles dedicated to corporal legal punishments, being outnumbered only by the monetary penalties. Nine of these represent the different modalities of capital punishment: *ajusticiar*, *arcabucear*, *crucificar*, *enrodar*, *executar*, *garrote*, *pena capital* (*s. v. pena*), *pena ordinaria* (*s. v. pena*), *poner en un palo* (*s. v. palo*). The social segregation is, once again, demonstrated in the lack of empathy and the severity of the sentences for ordinary people in contrast to the clemency shown to the noble and privileged classes. That can be grasped in the definition of the sub-article *castigo ò pena de azótes* which outlines this concept as punishment that causes infamy and regularly consists of 200 whip blows, and which is imposed on delinquents who

[16] These definitions of the canonical and the ecclesiastical law are given in the glossary by López Álvarez/Ortega Giménez (2010, p. 168).

[17] The tendency of preserving the archaic lexicon in the dictionaries of the Spanish Royal Academy has been demonstrated by Jiménez Ríos (2001), Ruhstaller (2002) and Freixas (2003), among others.

[18] Cf. Alzate Echeverri (2018) for concept of legal medicine in the 18th century.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

are not noble. However, the definition of the entry *azotar* that denotates the same concept, describes it as a penalty for "those delinquents that deserve such a punishment due to their crimes" indicating the changing perspective on the relation between the level of guilt and the punishment.[19]

3. The concept of judge in *Autoridades*

One of the defining factors of the legal dimension of the society of the Old Regime was the fragmentation of the juridical system.[20] Besides the royal jurisdiction and the local jurisdictions (such as municipal and seigniorial jurisdiction), special jurisdictions had been created in relation to different subject matters. These include jurisdictions, with their respective courts and officials, that dealt with domains such as trade (*alcalde alamin*),[21] religious matters (*juez conservador*),[22] military (*mariscal*), or universities (*juez del estudio*), but the jurisdictions that concerned very specific issues such as so-called *jurisdicción de la Mesta* (*juez entregador*) that used to resolve legal disputes of cattlemen.

A direct consequence of such a juridical system was the existence of numerous judicial, quasi-judicial, and advisory bodies, some royal and some regional, with overlapping jurisdictions and a perplexing hierarchical structure. A broad image of this phenomenon can be perceived in the lexicographic treatment of the modern concept of judge, i. e., a judicial official entrusted with the jurisdictional power to interpret the law, process, and resolve trials, as well as to execute the respective sentence.[23] This concept is treated within a total of 63 entries and subentries related to the different officials who, among their other duties, had the authority of presiding over different types of court proceedings. As many as 47 of these are lemmatized as either an *alcalde* ('mayor') or as a *juez* ('judge').

The office of an *alcalde* in the 18th century was similar to that of a modern one, with the addition of the duties of administrating justice.[24] This profession is, therefore, easily confused with that of a *juez*. The lexicographic entry in the *Autoridades* devoted to the noun *alcalde* defines it as "the person enjoying the dignity of judge, that administers justice in the town under their jurisdiction" (*s. v. alcalde*).[25] On the other side, the noun *juez* is defined as "the one who has authority and power to judge" (*s. v. juez*).[26] There are seventeen different subentries devoted to *alcalde* exercising different types of jurisdictions. Thirteen of these refer to the concept of a judge of ordinary jurisdiction (*alcalde de alzadas*, *alcalde de casa*, *alcalde de gradas*, *alcalde de hijosdalgo*, *alcalde de la hermandad*, *alcaldes del crimen*, *corte y rastro*, *alcalde mayor*, *alcalde mayor*, *alcalde ordinario*, *alcalde pedaneo*, *alcalde*, *alcaldes de hijosdalgo*) and four to judges of special jurisdictions (*alcalde alamin*, *alcalde de la mesta*,

[19] On the penal enlightenment in Spain, cf. Agüero/Lorente (2012).

[20] The diversity among the kingdoms that formed part of the Hispanic Monarchy —with Castile and Leon on the one side, and Aragon, Catalonia, and Valencia each in possession of their own legislation— led to a justice system that was fragmented on numerous levels.

[21] For all of the subentries related to the concept of *alcalde*, *s. v. alcalde*.

[22] For all of the subentries related to the concept of *juez*, *s. v. juez*.

[23] Definition of the noun "juez" ('judge') as given in a dictionary by Pallares (1986, p. 460).

[24] *DPEJ* (2016), *s. v. alcalde*, *desa*.

[25] "ALCALDE. s. m. La persona constituida en la Dignidád de Juez, para administrar justícia en el Pueblo en que tiene la jurisdicción […]" (*s. v. alcalde*).

[26] "JUEZ. s. m. El que tiene autoridad y poder para juzgar." (*s. v. juez*).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

alcalde de obras y bosques, *alcalde de sacas*, *alcalde mayor entregador*). Moreover, there is a total of thirteen entries treating different types of the concept of *juez*. Eight of these refer to the concept of the judge of ordinary jurisdiction (*jueces de competéncias*, *juez de enquesta*, *juez conservador*, *juez de commission*, *juez in curia*, *juez mayor de Vizcaya*, *juez supremo*, *juez*), and five to that of specialised jurisdictions (*juez escolástico*, *juez del estúdio*, *juez entregador*, *s. v. mariscal*). The definitions in these entries illustrate a thin line between the concept of *alcalde* and the concept of *juez* from the juridical point of view, such as following examples:

(4) ALCALDE DE CASA, CORTE, Y RASTRO. **Juez** que usa de Garnacha, y vara: tiene la jurisdición ordinária en la Corte, y cinco leguas en contorno: y para conocer de hurtos se extiende à veinte [...];

(5) ALCALDE DE LA MESTA. **Juez** nombrado por la quadrilla de Ganadéros, y aproba-do por el Concéjo, para conocer de los pléitos de pastóres [...];

(6) ALCALDE MAYOR. **Juez** de letras sin Garnácha, con jurisdición ordinária, aprobado por el Rey en su Consejo Real y Cámara de Castilla [...];

The role and the responsibilities of these officials vary widely across different jurisdictions and the lexicographical information introduced in *Autoridades* conveys an image of this variation. The degree of authority is often portrayed by two objects that symbolized the judicial power in the 18th century: *la garnácha* and *la vara*. The first one refers to the distinctive clothing that was used exclusively by the counsellors and the judges of the Real Audiencia[27] (*s. v. garnácha*), while the second one represents the cane used by certain officials as an emblem of their authority (*s. v. vara*). The dictionary often relies on these two symbols to illustrate the level of judicial power and the ranking among the numerous officials in charge of administering justice:

(7) ALCALDE DE GRADAS. [...] **Usan Garnácha, y vara**: tienen la jurisdición ordinária en su território, y forman sala para determinar las causas crimináles [...];

(8) ALCALDE DE OBRAS Y BOSQUES. [...] **Trahe Garnácha, y vara**; pero no la puede levantar en la Corte, sino solo en los bosques, y sítios de la casa del campo [...];

The perplexity of the hierarchical structure can be observed in the lemmatization and the definitions of the two different concepts as applied to the judge of noblemen:

(9) ALCALDES DE HIJOSDALGO. Se llaman los que resíden en las Chancillerías de Valladolíd y Granáda, donde forman sala con Escribános de Cámara [...] Conocen de los pléitos de hidalguía, y agrávios que se hacen à los Hidalgos [...] **Trahen Garnácha, pero no usan vara** [...];

(10) ALCALDE DE HIJOSDALGO. Se llama en los lugáres donde hai mitad de oficios el Alcalde ordinário, nombrado por el estado de los hijosdalgo. **Trahe vara, pero no Garnácha**. Se elíge todos los años, y es acto distintivo de nobleza [...].

At this point, one can perceive an additional aspect of the previously mentioned justice fragmentation, that relies on social stratification. While royal justice rested with the king and was in force across the entire territory of the monarchy, regional justice was often in the hands of the noblemen (*s. v. señorio*, *señor*) or the abbots of the monasteries that had the right to judge both civil and criminal matters (*s. v. vicariato*, *vicario*). Moreover, the nobles were assigned special jurisdiction, which provided their own judges (*alcalde de la hermandad*, *alcalde de hijosdalgo*, *alcaldes de hijosdalgo*) that adjudicated over legal disputes among them.

[27] A court in charge of administrating royal justice.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Apart from the two groups of nouns lemmatized as either *juez* or *alcalde*, there are sixteen additional entries that treat professions with capacities similar to those of judge of ordinary jurisdiction (*s. v. adelantado*, *alguacil*, *assistente*, *alamin*, *asses-sor*, *associado*, *auditor*, *baile*, *conjudice*, *corregidor*, *judicantes*, *justicia de aragón*,[28] *magistrado*, *magistrado*, *merino*, *ministro*, *oiidor*, *ordinario*, *pesquisidor*, *regente*, *señor*, *sequestro*, *tabla*, *visitadór*, *yuge*), as well as thirteen entries devoted to the officials of specialised jurisdictions (*auditor de la camara*, *auditor de rota*, *auditor del nuncio* (*s. v. auditor*), *contador mayor de cuentas*, *contadores de nombramiento* (*s. v. contador*), *datario*, *inquisidor general* (*s. v. inquisidor*), *inquisidor*, *inquisidores*, *provisor*, *veedor o juez del contraband* (s. v. *contrabando*), *vicario*). Nevertheless, in certain cases, such as the definition portraying the office of Hispanic royal judge known as *corregidor*, the definition focuses entirely on other, mostly administrative, functions and fails to mention the judicial dimension of these professions (s. v. *corregidor*).

A clear conclusion may be drawn from the foregoing: the lexicographic treatment of those officials whose duties were not only to resolve legal disputes, but also to represent "the royal persona, and judge, as the King himself, according to God's will on earth, the known truths, […] and according to what their conscience dictates, and they can exceed the laws" (Castillo de Bobadilla 1775, V, 3, 58)[29] overcomes a mere linguistic purpose and demonstrates an effort to detangle the complicated hierarchical structure, thereby reflecting the essence of the fragmentation of the legal system of that time.

4. Conclusion and future work

The above findings and discussion indicate the omnipresence of the extralinguistic factors in the definitions of the legal lexicon introduced in the first dictionary of the Spanish Royal Academy. The fact that the legal domain is sociologically, historically, and geographically limited, with its characteristics closely related to the context in question, reflected on the lexicographic treatment. Moreover, the research shows that, despite the heterogeneity of the employed texts, the first Spanish academics had a clear preference when it came to sources of legal lexicon. Finally, the extracted data reveals the significant variations in labelling of the lexical units related to the legal domain.

This research constitutes the initial steps of the project of the thesis. The thesis intends to contribute primarily to the study of the history of the reception of this technical lexicon within the model of Spanish academic lexicography, and thereby to the investigation of the history of lexicography, as well as to the knowledge of the semantic evolution of the legal lexicon.

References

Agüero, A./Lorente, M. (2012): Penal enlightenment in Spain: from Beccaria's reception to the first criminal code. In: Forum historiae iuris. https://forhistiur.net/2012-11-aguero-lorente/ (last access: 21-03-2022).

Alzate Echeverri, A. M. (2018): Reconocedores: médicos, empíricos y profanos en las decisiones judiciales. Nuevo Reino de Granada, siglo XVIII. In: Anuario Colombiano de Historia Social y de la Cultura 45 (1), pp. 47–78.

[28] S. *v. justicia.*

[29] "Representan la persona Real y como el Rey juzgan según Dios en la tierra, la verdad sabida […] y según les dicte su conciencia, y pueden exceder de las leyes" (Castillo de Bobadilla 1775, V, 3, 58).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Autoridades = Real Academia Española (1726–1739): Diccionario de Autoridades. https://apps2.rae.es/DA.html (last access: 21-03-2022).

Baldinger, K. (1985): Lengua y cultura: su relación en la lingüística histórica. In: Revista Española de Lingüística 15 (02), pp. 247–276.

Blanco Izquierdo, M.ª A./Clavería Nadal, G. (2019): Y así se dice…: los ejemplos y las notas de uso en los diccionarios académicos (1726–1852). In: Azorín, D./Clavería, G./Jiménez Ríos, E. (eds.): ELUA: El diccionario de la Academia y su tiempo: lexicografía, lengua y sociedad en la primera mitad del siglo XIX (5), Alicante, pp. 339–379.

Blecua, J. A. (2006): Principios del Diccionario de Autoridades, Real Academia Española. Madrid.

Cabré, M. T./Estopà, R./Lorente, M. (2011): Ideología y diccionarios especializados. In: San Vicente, Félix/Garriga, Cecilio/Lombardini, Hugo E. (eds.): Ideolex. Estudios de lexicografía e ideología. Monza, pp. 103–121.

Castillo De Bovadilla, J. (1704): Política para corregidores y señores de vassallos en tiempo de paz, y de guerra. Antwerp.

Desporte, A. (1998–1999): Les entrées non autorisées dans le Diccionario de Autoridades. In: Cahiers d'études hispaniques medievales (22), pp. 325–346.

DPEJ = Real Academia Española (2016): Diccionario panhispánico del español jurídico. https://dej.rae.es/ (last access: 21-03-2022).

Freixas Alás, M. (2003): Las autoridades en el primer Diccionario de la Real Academia Española. PhD tesis. Barcelona.

Freixas Alás, M. (2006): Los textos legales en el Diccionario de Autoridades. In: Clavería Nadal, G./Mancho, M. J. (eds.): Estudios de léxico y bases de datos, Universitat Autònoma de Barcelona. Barcelona, pp. 49–76.

Freixas Alás, M. (2010): Planta y método del Diccionario de Autoridades. Orígenes de la técnica lexicográfica de la Real Academia Española (1713–1739). A Coruña.

García-Gallo, A. (1971): La división de las competencias administrativas en España en la Edad Moderna. In: Actas del II Symposium de Historia de la Administración. Madrid, pp. 293–306.

González Alonso, B. (1970): El corregidor castellano (1348–1808). Madrid.

González Gilarraz, Mª del M. (1996): La administración de justicia ordinaria en la Edad Moderna en la Corona de Castilla: Procedimientos y tipos documentales. In: La investigación y las fuentes documentales de los Archivos. Guadalajara, pp. 485–498.

Gutiérrez Rodilla, B. M. (1994–1995): Construcción y fuentes utilizadas para los términos médicos en el Diccionario de Autoridades. In: Revista de lexicografía (1). A Coruña, pp. 149–162.

Henríquez Salido, M. d. C. (2010): Historia del léxico jurídico. Madrid.

Jiménez Ríos, E. (2001): Variación léxica y diccionario: los arcaísmos en el diccionario de la Academia. Madrid/Frankfurt a. M.

Lázaro Carreter, F. (1972): Cronica del Diccionario de autoridades (1713–1740). Real Academia Española. Madrid.

López Álvarez, A./Ortega Giménez, A. (2010): Glosario Jurídico Básico. Alicante.

López Díaz, M. (2006): La administración de la justicia señorial en el Antiguo Régimen. In: AHDE (76), pp. 559–583.

Mantecón Sancho, J. (2018): Pluralismo religioso, Estado y Derecho. Curso de Derecho Eclesiástico del Estado. Madrid.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Nielsen, S. (2015): Legal terms in general dictionaries of English: the civil procedure mystery. In: Lexikos (25), pp. 246–261.

Olivecrona, K. (1999): Lenguaje Jurídico y Realidad. Mexico City.

Pallares, E. (1986): Diccionario de Derecho 17a Procesal Civil. Mexico City.

Pérez Pascual, J. I. (2012): El léxico de especialidad. In: Léxico español actual (3), pp. 189–219.

Rodríguez Barcia, S. (2004): El léxico religioso en el Diccionario de Autoridades (1726–1739). In: Corrales Zumbado, Cristóbal José/Luis, Josefa Dorta/Nelsi Torres González, Antonia/Corbella Díaz, Dolores/del Mar Plaza Picón, Francisca (eds.): Nuevas aportaciones a la historiografía lingüística. Actas del IV Congreso Internacional de la SEHL, La Laguna, 22–25 October 2003. Madrid, pp. 1417–1426.

Ruhstaller Kuhne, S. (2002): Variantes léxicas en el Diccionario de Autoridades. Descripción lingüística y juicios normativos. In: Echenique Elizondo, M. T./Sanchez Mendez, J. (eds.): Actas del V Congreso Internacional de Historia de la Lengua Española (2), Valencia 31 January – 4 February 2000. Madrid, pp. 2321–2330.

Santayana Bustillo, L. (1979): Gobierno político de los pueblos de España, y el corregidor, alcalde y juez en ellos. Madrid.

Tiersma, P. (2008): The nature of legal language. In: Dimensions of forensic linguistics. Amsterdam, pp. 7–25.

Contact information

Marija Žarković
Universitat Autònoma de Barcelona
marija.zarkovic@e-campus.uab.cat

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

(Historical) Lexicology

EURALEX XX 22
XX EURALEX INTERNATIONAL CONGRESS
DICTIONARIES AND SOCIETY
12-16 July 2022, Mannheim, Germany

Ellert Thor Johannsson

OLD WORDS AND OBSOLETE MEANINGS IN MODERN ICELANDIC

Abstract This paper examines a certain subset of the vocabulary of Modern Icelandic, namely those words that are labelled as 'ancient' in the *Dictionary of Contemporary Icelandic* (DCI). The words were analysed and grouped into two main categories, 1) Words with only 'ancient' sense(s) and 2) words that have modern as well as an obsolete older sense. Several subgroups were identified as well as some lexical characteristics. The words in question were then analysed in two other sources, the *Dictionary of Old Norse Prose* (ONP) and the *Icelandic Gigaword Corpus* (IGC). The results show that the words belong to several semantic domains that reflect the types of texts that have survived until modern times. Most of the words are robustly attested in Old Norse sources, although there are a few exceptions. Large majority of the words can be found in Modern Icelandic texts, but to a varying degree. Limits of the corpus material makes it difficult to analyse some of the words. The result indicate that the words labelled 'ancient' can be divided into three main groups: a) words that are poorly attested and should perhaps not be included in the lexicographic description of Modern Icelandic; b) words that are likely to occur sometimes in Modern Icelandic; c) words that function as other inherited Old Norse words and perhaps do not require a special label or should have an additional sense in the DCI.

Keywords Modern Icelandic, Old Norse, historical lexicology

1. Introduction

The topic of this paper relates to the history of the vocabulary of Modern Icelandic. Old Norse/Icelandic has a continual text tradition from the 12th century until modern times and is structurally conservative. Most of the basic vocabulary of Modern Icelandic comes from Old Norse. The aim is to take a closer look at those words that specifically refer to the medieval language and have limited use but are recorded in lexicographic descriptions of the contemporary language. The paper is organized as follows: After the introductory section the sources for investigating old words and obsolete meanings are discussed. In section 3 there is an account of a survey of the words found in *The Dictionary of Contemporary Icelandic* (DCI) that have received diasystematic labelling as being 'ancient', i. e. belonging to the medieval language. I subsequently group these words into categories and subcategories, based on lexical criteria. I look at how these words are attested in the medieval sources for comparative purposes. I then look at how they are used in the contemporary language, based on results from modern corpora. Finally, there are some concluding remarks.

2. The sources for old words and obsolete meanings

What constitutes an old word, or an obsolete meaning is not very clear-cut. The approach taken in this paper is looking at the work of lexicographers and how they have chosen to label words in dictionaries. The caveat for this approach is that such labelling is not always consistent and is likely to include some judgment calls of individual editors. However, there are no obvious alternatives available.

Taking dictionary labelling as starting point we have a way to examine more closely how labelled words are used in other sources, such as reference works for the medieval language

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

as well as modern language corpus resources. For the investigation discussed in section 3 below I used the three resources that are further described in the following subsections.

2.1 The Dictionary of Contemporary Icelandic (DCI)

The source for the primary data is *Íslensk nútímamálsorðabók* or The Dictionary of Contemporary Icelandic (DCI). This is a recent online dictionary project established at the Department of Lexical Research and Lexicography at the Arni Magnusson Institute for Icelandic Studies in Reykjavík and edited by Halldóra Jónsdóttir and Þórdís Úlfarsdóttir (cf. DCI). This dictionary was launched in 2016 and contains around 56000 headwords covering the core vocabulary of modern Icelandic with new words and phrases added regularly as well as new examples of usage. Newly added material in the dictionary reflects the use in actual texts and is based on the Icelandic Gigaword Corpus (cf. Steingrímsson et al. 2018). In addition, the dictionary contains information on inflexion of words, by linking to another web-based resource: The database of Icelandic Morphology (Bjarnadóttir/Hlynsdóttir/Steingrímsson 2019).

DCI is a monolingual dictionary, but based on the multilingual dictionary ISLEX, which is also published and managed by the same institute. ISLEX is an online Icelandic dictionary between Icelandic and the other Nordic and Scandinavian languages, i. e., Danish, Swedish, Faroese, Norwegian Bokmål, Norwegian Nynorsk as well as Finnish. The lexicographic data generated by the ISLEX-project was subsequently used to create the new modern Icelandic monolingual dictionary (cf. Jónsdóttir/Úlfarsdóttir 2019). The dictionary accounts for the active vocabulary of Icelandic from around 1950 to the present day.

DCI includes some words that belong to earlier stages of the Icelandic language, as well as some words from Old Norse. These words are labelled in the dictionary database and are a part of a system of diasystematic labelling the dictionary employs to supplement the lexical information. These labels indicate specialized or restricted use of a particular word or meaning. The labels are considered a helpful tool for the user to get more detailed information about use of a word and include several linguistic registers. There are 24 such labels and they include: "informal", "formal", "old", "old-fashioned", "ancient", "poetic" and "pejorative" (the 24 labels are discussed in Jónsdóttir/Úlfarsdóttir 2019, p. 18). The use of this labelling is mostly based on the subjective evaluation and judgment of the dictionary editor when editing a dictionary entry, although some can be defined by more concrete criteria, e. g. poetic words. Some words can receive more than one label, e. g. "ancient" and "poetic" or "old-fashioned" and "informal".

As this investigation is into the vocabulary of the oldest recorded stage of the language and how it is reflected in a modern lexicographic description, I was primarily interested in words labelled 'old' or 'ancient'. I wanted to find out exactly how these categories are defined and what kind of words were included in each one.

When looking at the words with the labels 'old' vs. 'ancient' the difference between them is clearly visible. Words labelled 'old' are used in early modern Icelandic texts, many of them referring to elements of the old agricultural society of Iceland before the 20th century. An example is the word *héraðsskóli* 'district school' which refers to a special kind of school common in the 19th century but no longer operating. Another example is *hórdómsbrot* 'indecency offence', which is a legal term over an offence that no longer is punishable. Neither of these words are very old and are not found in Old Norse sources. Since I was more interest-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

ed in the oldest part of the vocabulary, I decided to limit this investigation to the words labelled *fornt* 'ancient'. These older words can easily be extracted from the dictionary and further analysed.

2.2 A Dictionary of Old Norse Prose (ONP)

For investigating the distribution and use of words in the medieval text sources I use the main lexicographic description of the language of medieval Iceland, namely *A Dictionary of Old Norse Prose* (ONP). This dictionary project is hosted at the University of Copenhagen and aims to semantically analyse an extensive collection of citations from all genres of Old Norse prose texts. This dictionary has been online since 2010 but was enhanced and relaunched in 2019 (ONP Online). ONP covers the period from the oldest preserved texts (from around 1150) up until the late Middle Ages (around 1540 for Icelandic). This dictionary seeks to render the lexicographic material as close to the original sources as possible giving a lot of attention to textual detail (cf. Johannsson/Battista 2014). The project has a complicated history as it has gone through the process of transitioning from print to a digital version in the middle of a multi-volume publication which has led to some internal discrepancies in the organization of the dictionary entries (cf. Johannsson/Battista 2016). The dictionary accounts for all preserved genres of prose texts, with large number of examples of word use and various supplementary secondary information about text types, manuscripts, provenience, and date of earliest use so it is well suited for studies into the vocabulary of the period and its semantic development.

The dictionary of Old Norse Prose is very expansive, and its citation collection includes 5–10% of all the texts produced in the medieval period. It is very likely that words labelled as "ancient" in a modern Icelandic resource would be found in this dictionary. One can assume that ONP has recorded the usage and some attestations of every word perceived to belong to the vocabulary of medieval Icelandic.

2.3 The Icelandic Gigaword Corpus (IGC)

The Icelandic Gigaword Corpus (IGC) is a large depository of Modern Icelandic texts and currently consists of over 1500 million running words. The corpus is tagged so each word contains information about a morphosyntactic tag and lemma. In addition to this each text is accompanied by bibliographic information. This corpus has been used in different language technology projects. More information about IGC and other related Icelandic corpora can be found Steingrímsson et al. (2018) and Steingrímsson/Barkarson (2020).

The IGC is a useful tool to investigate the distribution and use of particular words, such as the ones under scrutiny here, in Modern Icelandic texts. Since the corpus is tagged it can search for different inflectional forms of a particular headword. The IGC is therefore well suited for checking the prevalence of the words labelled as 'ancient' in the DCI to see how they are attested in Modern Icelandic text data.

3 The current investigation

The starting point for the current investigation is all the words in DCI that are labelled 'ancient'. The stated reason why certain words receive this label in the dictionary is that the editors wanted to include words that students can expect to come across when pursuing

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

their studies in Icelandic (cf. DIC). Since such words are not part of the active vocabulary of Modern Icelandic, they are labelled in this way to indicate to the users that these words are outside of the expected scope of the dictionary but could be encountered in some specialized context.

To obtain the data on these words I contacted the editors of the dictionary and asked them for a list of the relevant words along with all the information that is also part of the lexical description found in each entry in the dictionary. The information received allowed me to construct a small database with some of the key information about the words.

The DCI is a monolingual dictionary, and the definition of the word is usually rendered by giving a short explanation. This is clearly visible in Figure 1, which contains a typical entry from the DCI.

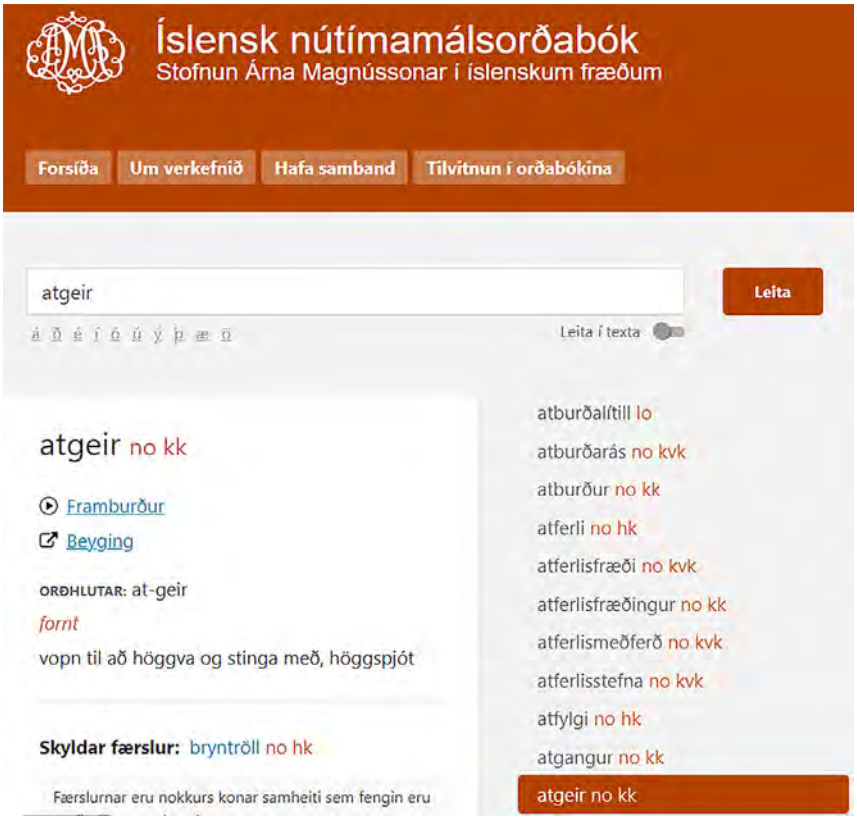


Fig 1: The headword *atgeir*. The word is defined as 'weapon for slaying or stabbing, bill'. The label *fornt* 'ancient' is shown in light brown colour

In total there are 65 items that are labelled as 'ancient' in DCI. The first step in the investigation was to take a closer look at these words to see if they had any identifiable common features, or if any patterns could be observed.

3.1 The survey

A survey of the 'ancient' words revealed that they can roughly be divided into two main categories. The first category includes words or lexical items that are part of the medieval linguistic registry and do not have different modern meaning. All in all, there are 49 items that belong to this first category. The second category are words that also exist in Modern

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Icelandic but have a different meaning than in the medieval language, in most cases having acquired a new sense that is also listed in the DCI. All in all, there are 16 items that belong to this second category.

The words from the first category are listed in table 1 in alphabetical order along with some morphological information and translated dictionary definitions as well as subgroup (sub) (see section 3.2 and 3.3 below).

Word	**Gram.**	**Mor**	**Definition**	**Sub**
atgeir	n m	c2	a 'casting spear'	bat
ben	n f/n	b	wound	bat
bifröst	n f	c2	rainbow	rel
bleyði	n f	b	cowardness	soc
bolöxi	n f	c2	axe for chopping wood	bat
brullaup	n n	c	wedding	soc
bryntröll	n n	c2	long shaft axe with a spear point	bat
dæll	adj	b	friendly, comfortable	alt
eiðsvari	n m	c2	compurgator, sworn follower	leg
firrast	v	b	avoid (s-t)	spec
fjörbaugsgarður	n m	c3	legal punishment of lesser outlawry	leg
fjörráð	n npl	c2	plan to kill someone	bat
forvitri	adj	c2	1. clairvoyant, 2. very wise	rel
geirlaukur	n m	c2	garlic	alt
goðorðsmaður	n m	c3	a chief in medieval Iceland	leg
gás	n f	b	goose	alt
harðhugaður	adj	c2	who is rough and harsh	bat
heimskringla	n f	c2	earth/world	soc
hjálmvölur	n m	c2	tiller (on a boat)	naut
hrjóða	v	b	drive people of a ship (in battle)	bat
híð	n n	b	see *híði* animal den/lair	alt
húðfat	n n	c2	sleeping bag from sheep skin	soc
hörgur	n m	b	1. pagan temple/shrine, 2. pile of rocks	rel
jarteikn	n f	c2	omen	rel
jarðarmen	n n	c2	turf, sod	rel
knör	n m	b	ship/boat	naut
knörr	n m	b	ship/boat	naut
krosskirkja	n f	c2	cross shaped church	rel
liðsbón	n f	c2	asking for help	bat
log	n n	b	light, fire	alt
menntur	adj	b	be educated	soc

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Word	Gram.	Mor	Definition	Sub
Mikligarður	n m	c2	the capital of East Roman empire, Byzantium	soc
mundlaug	n f	c2	basin for washing hands	alt
nátt	n f	b	night	alt
prímsigning	n f	c2	primary baptism of pagans	rel
rann	n n	b	house, home	alt
rögn	n npl	b	pagan gods	rel
serða	v	b	penetrate (sexually)	soc
sjálfdæmi	n n	c2	self-judgment	leg
skógarmaður	n m	c2	outlaw	leg
skóggangsmaður	n m	c3	outlaw	leg
skóggangur	n m	c2	outlawry	leg
strandhögg	n n	c2	foray 'strandraid'	bat
sýr	n f	b	sow (i. e. female pig)	alt
valköstur	n m	c2	pile of dead bodies from battle	bat
vígur	adj	b	capable of battle	bat
vöttur	n m	b	mitten	soc
öndvegissúla	n f	c2	one of two pillars on each side of a high seat	soc
örendi	n n	b	1. to run out of breath / give up 2. to die	spec

Table 1: All the words in category 1 with grammatical mark-up and morphological information, n= noun, n= neuter, f=feminine, m=masculine, pl=plural, adj= adjective, v= verb, b=basic (uncompounded), c2=compound of two element, c3=compound of three elements

The words from the second category are listed in table 2 in alphabetical order along with some morphological information, translation of the dictionary definition, the old (obsolete) definition as well as subgroup (see section 3.2. below).

Word	Gram.	Mor	New sense	Old sense	Sub
aðsókn	n f	c2	attendance	attack	bat
bjóða	v	b	offer	order something to someone	soc
blóðrás	n f	c2	blood circulation	blood flowing, loss of blood	bat
drottinn	n m	b	god, christ	master, lord	soc
friðland	n n	c2	sanctuary	safe place	soc
ljósta	v	b	strike	strike down in battle	bat
lyfting	n f	b	lifting up something	steering platform at the back of ship	naut
skeina(st)	v	b	wipe a childs behind	be wounded	bat
skipti	n npl	b	exchange	interaction	soc
skör	n f	b	brink	footrest	spec
skrælingi	n m	b	offensive: barbarian	native of North-America	soc
skuggsjá	n f	c2	'the glass of time'	mirror	spec

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Word	Gram.	Mor	New sense	Old sense	Sub
snekkja	n f	b	yacht	long-ship	naut
staðfestast	v	c2	confirm	take up residence	soc
völva	n f	b	prophetess	sibyl	rel

Table 2: All the words in category 2 with grammatical mark-up and morphological information, see table 1 above for abbreviations

In addition to these words, there are 76 words that receive the label 'poetic' and some of those appear to be Old Norse as well. Since the ONP dictionary does not include the poetic language, these words have been left out of the discussion here except for the three words that are both labelled 'ancient' and 'poetic': *bifröst* 'rainbow, bridge', *knörr* 'ship' and *rann* 'house/home'.

3.2 Category 1

In this subsection I will look closer at this category to try to identify its characteristic features. This is the larger group of words and can be divided into several subgroups mostly based on semantic domains, but also other criteria (last column in table 1). Some of the words are clearly definable as having a particular meaning, such as weapons like *atgeir*, (cf. Fig. 1), whereas other are less concrete and often have a more wide-ranging meaning, such as *liðsbón* 'asking for assistance/troops'. I was able to identify seven relatively clear subgroups and assigned every word to one of those.

3.2.1 Battle words (11)

These are words that refer to a battle or fighting and include weapons as well as other terms relating to violent confrontations: *atgeir* 'a casting spear', *ben* 'wound', *bolöxi* 'a kind of axe', *brynträll* 'long shaft axe with a spear point', *fjörráð* 'plan to kill someone', *harðhugaður* 'who is rough and harsh', *hrjóða* 'drive people of a ship (in battle)', *liðsbón* 'asking for assistance/troops', *strandhögg* 'military loot', *valköstur* 'pile of dead bodies from battle' *vígur* 'capable of battle'.

3.2.2 Legal words (7)

These are the words relating to legal status, punishment, ownership and similar things having to do with official positions: *eiðsvari* 'compurgator, sworn follower', *fjörbaugsgarður* 'legal punishment of lesser outlawry', *goðorðsmaður* 'chief in medieval Iceland', *sjálfdæmi* 'self-judgment', *skógarmaður* 'outlaw', *skóggangsmaður* 'outlaw', *skóggangur* 'outlawry'.

3.2.3 Religious words (9)

These are words relating to religious affairs, both pre-Christian and Christian: *bifröst* 'rainbow', *jarteikn* 'omen', *jarðarmen* 'long elavated turf under which men would perform ritual to become blood brothers', *krosskirkja* 'cross shaped church', *prímsigning* 'primary baptism of pagans', *rögn* 'pagan gods'. Two of the words have two 'ancient' senses. The adjective *forvitri* refers to someone who can predict the future but also someone who is 'very wise'. The word *hörgur* is commonly used in Old Norse to refer to a 'pagan temple/shrine', but can also refer more generally to 'rocky piles'. The word for rainbow *bifröst* is in this subgroup

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

as the rainbow functioned as a bridge to the realm of the gods in the Pre-Christian religion of the Nordic countries.

3.2.4 Societal words (9)

The words are: *bleyði* 'cowardness', *Heimskringla* 'earth/world', *húðfat* 'sleeping bag from sheep skin', *menntur* 'who is educated'; *Mikligarður* 'old name for the capital of East Roman empire, Byzantium', *serða* 'penetrate (sexually)', *vöttur* 'mitten', *öndvegissúla* 'one of two pillars on each side of a high seat', *brullaup* 'wedding'. The words cover a broad range of meaning.

3.2.5 Nautical words (2)

These belong to the realm of navigation and sailing: *hjálmvölur* 'tiller (on a boat)' and *knör* 'ship/boat' also spelled *knörr*. These two variants of the same word are counted as two lexical items even though they are essentially the same word, referring to a particular kind of ship, popular with the Vikings. The variations can be explained as Modern Icelandic spelling vs. Old Norse spelling, but both are known in Modern Icelandic.

3.2.6 Alternative word forms (9)

This subgroup is not based on semantics, but contains words that share the common feature of having some morphological connection to a prevalent modern Icelandic word, while having fallen out of use themselves. The types of words are quite diverse: The adjective *dæll* 'friendly, comfortable', is probably listed to explain modern Icelandic antonym *ódæll* 'naughty', which is widely used. The modern Icelandic equivalent of *gás* 'goose' is *gæs*, a backformation of the plural *gæsir*. The word *nátt* 'night' is an earlier sideform of the common word *nótt* and shares some inflectional forms with it. The word *rann* 'house, home' is known from certain fixed expression and the verb *rannsaka*, which is cognate to English *ransack*. The word *sýr* 'sow' like common Icelandic *kýr* 'cow' has an unusual inflectional pattern. The word *geirlaukur* 'garlic' shows the same word formation as English *garlic* and slightly different from the common Icelandic word *hvítlaukur*. The word *log* 'light, fire' is very similar to its Modern Icelandic equivalent *logi*, The word *mundlaug* 'basin for washing hands' is a compound noun of the very common Modern word *laug* 'pool' and *mund*, which is an old word for hand, known in certain sayings and *handlaug* is well attested in Modern Icelandic. The final word in this group, *híð* 'animal den/lair', has a common modern equivalent, *híði*, which has the same meaning.

3.2.7 Exceptional words (2)

This subgroup includes two words that do not fit into any of the other groups mentioned. The verb *firrast* 'avoid s-t' is somewhat problematic as has a rather wide-ranging meaning and is frequently confused with the verb *fyrtast* 'be offended' in Modern Icelandic. Perhaps it should be counted with the words in 3.2.6 although the similarities between these verbs are superficial. The word *örendi* is only used in a fixed expression *þrjóta örendið* 'to run out of breath, give up' which is attested in Old Norse. In DCI it has a second 'ancient' sense referring to the last breath before death.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

3.3 Category 2

In this subsection I will look closer at this category to try to identify its characteristic features. This category contains words that exist in Modern Icelandic but have both an 'ancient' sense and a modern one listed in the DCI. The words can be grouped by semantic domain and other criteria, although there are fewer subgroups represented than in category 1 (cf. last column in table 2).

3.3.1 Battle words (5)

The word *blóðrás* is the common Modern Icelandic word used for 'blood circulation'. In Old Norse the meaning is quite different as it means 'blood flowing' or 'loss of blood' and usually in the context of a violent event. The verb *ljósta* means 'strike' in Modern Icelandic and usually used about lightning, but in Old Norse it has the meaning of 'striking someone down in battle'. The verb *skeina* in Modern Icelandic is used about 'wiping a child's behind' after a toilet visit. In Old Norse the meaning is quite different or 'to be wounded'. The middle voice form of the same verb is also a headword in the dictionary, *skeinast* 'to be wounded'. The contemporary meaning of *aðsókn* is 'attendance', like to a concert or some other event. In Old Norse the word overwhelmingly means 'attack', but there are a handful of examples from late sources where the word has acquired the modern meaning.

3.3.2 Societal words (6)

The Modern Icelandic *friðland* means 'sanctuary', e. g. for wild animals or birds. The meaning in Old Norse is broader and seems to be more like 'safe place'. The verb *bjóða* 'offer' or 'invite' is a very common verb in Modern Icelandic but in Old Norse this word also has another meaning that is now lost in the contemporary language, which is 'to order someone to do something'. The word *drottinn* is very common in Christian context and refers to both God and Christ as 'lord'. In Old Norse the meaning was often more worldly and meant something like 'master', although the word could also refer to Christian deities. The word *skipti* means 'to exchange something' or 'switch something out' in Modern Icelandic, but in Old Norse it could have the meaning 'interaction or communication between people, although this word has several meanings already in medieval times. The word *skrælingi* is pejorative or offensive in Modern Icelandic and means something like 'barbarian' or a 'primitive human'. In Old Norse this term occurs in sagas about the Viking explorations in North America and refers to the inhabitants of the western continent, native Americans and Inuits. The verb *staðfestast* means 'to be confirmed' in Modern Icelandic, but in the medieval language it could also mean 'to take up residence', but this meaning is lost in contemporary Icelandic.

3.3.3 Religious words (1)

A *völva* 'sibyl' is a soothsayer in medieval sources and in Modern Icelandic it is used to refer to a prophetess that prominently figures in certain tabloid newspapers every year to make a prediction for the new year.

3.3.4 Nautical words (2)

The word *lyfting* can have the meaning 'lifting something' in Modern Icelandic whereas in Old Norse it has a more concrete meaning as 'a steering platform at the back of ship'. The Modern Icelandic word for 'yacht' is *snekkja* and in Old Norse this word refers to a particular kind of ship called *longship*.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

3.3.5 Exceptional words (2)

There are three words that fall a bit outside the groups accounted for so far. These are *skör* which in Modern Icelandic means 'brink', but in Old Norse has commonly a more concrete meaning of 'footrest'. The word *skuggsjá* is somewhat unique in the sense that it has acquired a secondary restricted meaning in Modern Icelandic as 'filter, the glass of time' and lost its Old Norse more concrete meaning as 'mirror'.

3.4 The results of the survey

The survey of the words labelled 'ancient' reveals some of their characteristic features. The division into subgroups is mostly based on subject fields that seem to be somewhat clearly definable. It is likely that the prevalence of the subgroups observed is that these groups reflect the types of texts that have been preserved from the medieval period, which are primarily sagas, i. e. stories of heroic achievements, religious texts and legal material. The subject matters of these texts often include concepts and concrete things related to the structure of the medieval society and frequently involve travel, conflict and legal disputes.

The two categories identified show some difference in the number of subgroups, as no category 2 words can be grouped as belonging to the legal domain or the alternative wordform subgroupp. The reason might simply be lack of data as the category 2 words are somewhat fewer in number. However, it could also indicate that some subgroups contain words that are more susceptible for acquiring a new meaning or changing meaning than others. This would have to be investigated further.

The words that fall under the heading 3.2.6 are somewhat different from the rest as they are not defined by semantic criteria, but rather formal characteristics. Some of them could also be grouped according to their semantic domain, even though most of them do not fit into any of the other subcategories. Here a preference has been made to classify these words in relation to their modern Icelandic variants or related words, as this would be the primary reason for their inclusion in the DIC. Their semantic domain is less relevant.

The survey of the words under scrutiny here illustrates the features of the words that belong to an earlier period that are likely to be relevant in modern context. The survey also reveals the types of words and meanings that dictionary editors consider appropriate for receiving the diasystematic label 'ancient' when describing the vocabulary of a modern language.

3.5 Attestation in Old Norse

The next step in the investigation was to look at the attestation of all the words in the surviving prose text corpus of Old Norse as it is recorded in the ONP dictionary. The purpose of this is twofold: Firstly, to check if the information from this source is in line with the information found in the DCI. Secondly to find out how well these words are attested, i. e., how many examples there are of each word in the documented vocabulary of medieval prose. Of course, the documented vocabulary does not reflect the prevalence of these words in Old Norse, as many registers of the language are not written down and the surviving texts do not necessarily give an accurate picture of word use. Furthermore, even if we accept these caveats the information is not always completely straightforward as the ONP dictionary is incomplete and many of the headwords have not yet been edited in a structured

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

dictionary entry. Nonetheless, all the citations are accessible in ONP, so the user can get some idea about the use and meaning of any word, even the ones that are unedited.

The results are shown in table 3. When there were multiple senses in the semantic structure, I was usually able to assign a word to a particular sense, and in such cases the total number of citations is shown in parenthesis (i. e. all senses).

Cat	Headword	ONP	IGC
1	atgeir	45	195/149
1	ben	58	211/26*
1	bifröst	2	17902/6
1	bleyði	20	3
1	bolöxi	50	5
1	brullaup	196	24
1	bryntröll	14	15
1	dæll	44	1960/8*
1	eiðsvari	20	0
1	firrast	55	94
1	fjörbaugsgarður	10	32
1	fjörráð	68	46
1	forvitri	28	20
1	gás	23	139/11*
1	geirlaukur	1	15
1	goðorðsmaður	26	267
1	harðhugaður	3	0
1	heimskringla	17	1514/182
1	híð	9	1
1	hjálmvölur	22	0
1	hrjóða	38	12
1	húðfat	49	3
1	hörgur	22	138
1	jarðarmen	12	192/190
1	jarteikn	264	70/52
1	knör	42	89/41
1	knörr	>knör	775/266
1	krosskirkja	7	102/41
1	liðsbón	8	31
1	log	33	N/A
1	menntur	27	172/123
1	Mikligarður	24	1300
1	mundlaug	52	13
1	nátt	284	277/190*
1	prímsigning	7	6
1	rann	4	N/A
1	rögn	2	175/15*
1	serða	8	72/64*
1	sjálfdæmi	106	1563
1	skógarmaður	52	282/129
1	skóggangsmaður	2	17/16
1	skóggangur	7	41
1	strandhögg	44	512
1	sýr	5(12)	243/15*
1	valköstur	34	26
1	vígur	61	124/38*
1	vöttur	7	258/14*
1	öndvegissúla	17	203
1	örendi	5	330
2	aðsókn	37	21996
2	bjóða	799*	622172
2	blóðrás	93	1488
2	drottinn	98(188)	18475
2	friðland	34	6522
2	ljósta	329	5693
2	lyfting	34	3579
2	skeina	22	311
2	skeinast	57	16
2	skipti	40(135)	247,605
2	skrælingi	10	151
2	skuggsjá	3	168
2	skör	18(42)	7494
2	snekkja	29	3169
2	staðfestast	56	1678
2	völva	22	1437

Table 3: The number of attestations for each word in A Dictionary of Old Norse Prose (ONP) and the Icelandic Gigaword Corpus (IGC)

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Most of the words are quite common, with more than 10 instances recorded in the ONP database. It is of course likely that common words are more prone to show up in Modern Icelandic context, either in texts discussing the medieval society or citing certain passages from older literature. The verb *skeina*/*skeinast* can be separated here according to formal criteria, whereas *knör*/*knörr* is reflective of the same word.

3.6 Attestation in Modern Icelandic

The final step of the investigation was to look at the attestation of these words in Modern Icelandic. This was done by looking at examples from the Icelandic Gigaword Corpus (IGC) (Steingrímsson et al. 2018). This source can provide detailed information about the prevalence of individual words based on structural criteria, but not semantic criteria. Therefore, analysis of category 2 words is very difficult, except where the examples are relatively few.

The results of the analysis are shown in table 3 above. The numbers indicate the absolute number of occurrences of the word before and after filtering separated by dash if there was a difference. I tried to filter out proper names and obvious errors if it was manageable (manual filtering indicated by an asterix *). There were examples of almost all of the words identified as 'ancient' in the DCI, but most of them are relatively rare. In what follows I will discuss the attestations of selected individual words.

Three words did not come up at all when searching the IGC. These are the nouns *eiðsvari*, *hjálmvölur* and the adjective *harðhugaður*, which is also rather rare in ONPp. It is unlikely that these words show up in dictionary searches as they are uncommon and doubtful whether they should be included in the contemporary description of Icelandic vocabulary. The same is to say about the word *híð*, which only occurs once, in a poem.

There were also two words which showed up in the corpus, but the actual examples were hard to find because of noise in the data. The word *log* gets confused with the English word *log* and *rann* is mostly limited to fixed expressions *í eigin ranni* 'in own home', *í sama ranni* 'in the same home' when it is not confused with other lemmas. For these words I was not able to count the examples in any meaningful way.

Other rare words are *bleyði*, *húðfat* and *bolöxi*, which occur only in handful of examples. The word *prímsigning* is also rare and only occurs when discussing medieval society.

Some words are well attested as proper names, but when those are filtered out not many examples remain. This is especially noticeable for *Bifröst*, which is a name of an Icelandic educational institution as well as a placename and well attested as such. There are only a few examples of the word being used as a common noun in the meaning 'rainbow' or 'bridge' and they are almost all from poetry (and also labelled as such in DCI).

The use of the older form of the word for goose, *gás*, occurs mainly in fixed expression *gjalda gagl fyrir gás*. Most of the other examples in the corpus that are filtered out are of the plural *gæsir*, which is the regular plural of *gæs*, the Modern equivalent and of the placename *Gásar*.

The adjective *dæll* is affected by much noise in the data as many of the examples are erroneously tagged as participle forms of the verb *dæla* 'pump'. The actual adjectival forms turned out to be much fewer. Another word that was difficult to analyse because of similar problems was *vöttur* which got confused with the word *vatt* 'watt', a measuring unit for

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

electric power. The actual examples of this word turned out to be rather few and only referring to old handiwork and museum pieces.

Other words were also not as prevalent as the initial search indicated and required some manual analysis. The examples of the regular noun *ben* frequently turned out to be abbreviations of the common name *Benediktsson,* but those were not counted. More scrutiny of the search results for some other words resulted in large reduction of examples, such as in the case of *rögn* and *sýr* which also had proportionally many erroneous examples and the same is to say about the adjective *vígur.*

The numbers of examples for *skógarmaður, krosskirkja, heimskringla* are not as many as appear at first glance, as they can also refer to proper names. Even when they have been filtered out the word *skógarmaður* in most cases refers to a member of a YMC club and not an outlaw. This restricted meaning is not registered in DCI. *Krosskirkja* refers to a shape of a church that was common in ancient and medieval times and is mostly found in such context. *Heimskringla* is a famous literary work, but the word is also sometimes used in Modern Icelandic as a very formal way of referring to the earth.

Some words that clearly refer to medieval societal and religious phenomena are *fjörbaugsgarður*, *skóggangur*, *brullaup*, *hörgur*, *öndvegissúla* and *knör(r).* The last one is not predominantly poetic although it is labelled as such in DCI. Other words that are found in texts discussing medieval matters seem often to be used as Modern Icelandic words, such as *liðsbón*, *fjörráð* and *jarteikn*, mostly by politicians. The same is to say for the phrase *ganga(st) undir jarðarmen.*

The word *strandhögg* has many examples, has acquired a more figurative meaning and seems to be used as a fully functioning Modern Icelandic word meaning 'make an impression/dent' in sentences like *erlendir bankar munu gera hér strandhögg* 'foreign banks will make a dent in the market here'.

The words with the highest frequency in the ICG are *Mikligarður* and *sjálfdæmi.* The prevalence of those words can be explained as *Mikligarður* is a name of a supermarket which is mentioned more often than Byzantium. The word *sjálfdæmi* is quite frequent, especially in parliamentary speeches, and seems to mean the same as in Old Norse, i. e. the act of judging oneself or deciding for oneself how to act, e. g. *framselja sveitarfélögum sjálfdæmi um upptöku gjaldsins* 'grant power to the municipalites to decide for themselves whether to charge this fee'.

In general the words in category two are much more frequent as the more common meaning of those words in contemporary language, more often than not, is quite different from the one in the medieval language and impossible to find the relevant examples if they exist. The main exceptions are *skrælingi* and *skuggsjá* and they often refer to medieval context.

4. Conclusions

The current investigation into this limited part of the vocabulary of Modern Icelandic has revealed several characteristics of the words labelled 'ancient' in the Dictionary of Contemporary Icelandic and their distribution in old and new source.

The survey of the words revealed how they can be divided into two main categories The characteristic features of several subcategories were identified as well as some factors that could help explain the observed distribution.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Further comparison with Old Norse lexicographic data revealed that most of the words in question are well attested common Old Norse words, although there are some noticeable exceptions.

The analysis of modern Icelandic corpus data shows that we can divide the words into three distributional groups. The first group contains a few rare words, that are not part of the active vocabulary of Modern Icelandic and would most likely not be encountered by students. Such words could be omitted from a contemporary lexicographic description without drastic consequences. The second group is the largest one and contains most of the words that are well attested in Old Norse sources and frequently referred to in Modern Icelandic contemporary culture. It is therefore helpful for learners of Icelandic and other dictionary users to be able to look up the meaning of these words when they encounter them. The third group of words, which is also not big, are words that should perhaps not be labelled as 'ancient' or should receive an additional sense in the DCI, as they are fully functional lexical entities in Modern Icelandic and are used as any other contemporary Icelandic words.

References

Bjarnadóttir, K./Hlynsdóttir, K. I./Steingrímsson, S. (2019): DIM: "The Database of Icelandic Morphology". In: Proceedings of the 22nd Nordic Conference on Computational Linguistics, NoDaLiDa 2019. Turku, pp. 146–154.

DCI = Íslensk nútímamálsorðabók (The dictionary of contemporary Icelandic). Úlfarsdóttir, Þ./Jónsdóttir, H (eds.): Reykjavík.http://islenskordabok.arnastofnun.is (last access: 24-03-2022).

ISLEX = Úlfarsdóttir, Þórdís (chief editor.). Reykjavík. http://islex.is (last access: 24-03-2022).

Johannsson, E. T./Battista, S. (2014): A dictionary of Old Norse prose and its users – paper vs. web-based edition. In: Abel, A. et al. (eds.): Proceedings of the XVI EURALEX International Congress: The User in Focus, 15–19 July 2014. Bolzano/Bozen, pp. 169–179.

Johannsson, E. T./Battista, S. (2016): Editing and presenting complex source material in an online dictionary: the case of ONP. In: Margalitadze, Tinatin/Meladze, Georg (eds.): Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity, 6–10 September 2016. Tbilisi, pp. 117–128.

Jónsdóttir, H./Úlfarsdóttir, Þ. (2019): Íslensk nútímamálsorðabók. In: Orð og tunga 21, pp. 1–26. https://doi.org/10.33112/ordogtunga.21.2.

ONP Online = Ordbog over det norrøne prosasprog/A Dictionary of Old Norse Online. http://onp.ku.dk (last access: 24-03-2022).

Steingrímsson, S./Helgadóttir, S./Rögnvaldsson, E./Barkarson, S./Guðnason, J. (2018): Risamálheild: a very large Icelandic text corpus. In: Proceedings of LREC 2018. Myazaki, Japan, pp. 4361–4366.

Steingrímsson, S./Barkarson, S. (2020): Icelandic Gigaword Corpus 1 (IGC1) – version 20.05, CLARIN-IS. http://hdl.handle.net/20.500.12537/41 (last access: 24-03-2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Contact information

Ellert Thor Johannsson
The Arni Magnusson Institute for Icelandic Studies
etj@hi.is

Acknowledgements

I would like to thank Þórdís Úlfarsdóttir, one of the editors of the Dictionary of Contemporary Icelandic, for providing me with the data discussed in this article. I would also like to thank the anonymous reviewers who brought to light various problematic issues, which I have tried to resolve in this final version.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Pius ten Hacken/Renáta Panocová

THE ETYMOLOGY OF INTERNATIONALISMS

Evidence from German and Slovak

Abstract In the etymological information for a word in a dictionary, the first question to be answered is whether the word is a borrowing or the result of word formation. Here, we consider this question for internationalisms ending in *-ation* in German and in *-ácia* in Slovak. In German, *-ation* is a suffix that attaches to verbs in *-ieren*. For these verbs, it is in competition with *-ung*. In Slovak, *-ácia* is a suffix that attaches to bases of Latin or Greek origin. The corresponding verbs are often backformations. Most Slovak verbs also have a nominalization in *-nie*. In order to investigate to what extent the nouns in *-ation* or *-ácia* are borrowings or derived from the corresponding verbs in German and Slovak, we took a random sample of English nouns in *-ation* for which OED gives a corresponding verb. For this sample, we checked whether the cognate noun in *-ation* or *-ácia* is attested in standard dictionaries and in corpora. Then we did the same for the corresponding verbs and the nouns in *-ung* or *-nie*. Finally, we checked the frequency of these words in DeReKo for German and SNK for Slovak. On this basis, we found evidence that *-ation* in German has a slightly different status to *-ácia* in Slovak. This status affects the relationship to the corresponding verbs and to the nouns in *-ung* or *-nie*. Such generalizations are important as background information for specifying etymological information in dictionaries, especially for languages where first attestations dates are not readily available.

Keywords Borrowing; word formation; reanalysis

1. The suffixes *-ation* in German and *-ácia* in Slovak

In providing etymological information in dictionaries, a central question is whether the word in question is a borrowing or the result of word formation. Here, we will study this question for internationalisms that are marked by the suffix illustrated in (1).

(1) a. communication
b. Kommunikation
c. komunikácia

Internationalisms, as discussed for instance by Braun (1990) and Waszakowa (2003), are words that appear in different languages from different language families, as (1a) from English and French, (1b) from German, and (1c) from Slovak. The suffix in (1), which appears as *-ation* in English, French and German, and as *-ácia* in Slovak, can be traced back to Latin. However, as ten Hacken/Panocová (2022) show, it is not itself a suffix in Latin.

The reason for studying internationalisms such as (1) is that their fairly recent emergence gives the opportunity to study the question of whether they are the result of borrowing or word formation on the basis of documented sources. In English, OED (2000–2022) gives important information by dating the examples. First attestation dates can be used to support the hypothesis of borrowing or word formation as their source, as illustrated by ten Hacken/Panocová (2022). Here, we will focus on German and Slovak, two languages for which resources in the form of large corpora and scholarly dictionaries are available, but where first attestation dates of words have not been systematically documented.

The suffix *-ation* in German is a nominalizing suffix attaching to verbs with an infinitive in *-ieren* (cf. Fleischer/Barz 2012, pp. 242 f.). For these verbs, it is in competition with *-ung*. The

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

suffix *-ung* has a much broader range of application, including also verbs with a different infinitive ending, e.g. *Leistung* ('performance') from *leisten* (cf. Fleischer/Barz 2012, pp. 225–230). In addition, all verbs in German have a nominalized infinitive. Whereas nominalized infinitives generally only have a process reading, nominalizations in *-ation* and *-ung* can also have further readings. Thus, (1b) can refer to the transmission of information, but also to the information that is transmitted. Stem conversion, e.g. *Vergleich* ('comparison') from *vergleichen* ('compare'), does not occur for verbs in *-ieren*, so that it does not constitute a competition for the suffix *-ation*.

In Slovak, *-ácia* is the suffix corresponding to *-ation*. In the Slavic linguistic tradition, it is labelled as an international suffix. It is understood that internationalisms containing it are of Latin or Greek origin and they are sometimes called neologisms of Neolatin descent (Buzássyová 1992, p. 89; Horecký et al. 1989, pp. 130–132). For (almost) all Slovak verbs, it is possible to form a noun in *-nie*. Thus, for nouns in *-ácia* with a verb in *ovať*, there is a systematic synonymy between the nouns in *-ácia* and in *-nie*.

Both in German and in Slovak, many nouns of the type in (1) are borrowings. From the second half of the 20th century, these borrowings are usually from English. Earlier borrowings in German are often from French, whereas in Slovak, Latin often served as the source.

2. Data collection

For our study, we collected German nouns in *-ation* and Slovak nouns in *-ácia*. As our purpose was to compare the two languages in this respect, we did not use German or Slovak as the starting point for our collection of data, but we used a third language. We chose English as a starting point, because the documentation available in OED (2000–2022) provides detailed information and possibilities of retrieval that are not available for many other languages.

In OED, it is possible to retrieve all nouns in *-ation*. For these nouns, we checked whether a corresponding verb is recorded in OED. For our pilot study, we took a randomized sample of 200 nouns in *-ation* with a corresponding verb. For these nouns, we asked native speakers of German and Slovak to translate them and look in particular for cognates in *-ation* and *-ácia*. The attestation and frequency of these nouns was then verified in monolingual dictionaries and corpora. For German, we used Duden (2022) and DeReKo (W corpus). For Slovak, the main dictionary is KSSJ (2003) and the corpus is SNK. In addition, we recorded the attestation and the frequency of the corresponding verb and the competing derivations, i.e. the one with *-ung* in German and with *-nie* in Slovak. On the basis of English *organization*, the German triple in (2) and the Slovak one in (3) were found.

(2) a. Organisation
b. organisieren
c. Organisierung

(3) a. organizácia
b. organizovať
c. organizovanie

An important condition recorded in the sample is that of cognate nouns. For *evaporation*, the most common translation in German is *Verdunstung* and in Slovak *vyparovanie*. In such cases, these nouns and their corresponding verbs were recorded in our sample, but we also considered whether the cognates *Evaporation* in German and *evaporácia* in Slovak occurred

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

in the dictionaries and corpora. In this case, both cognates are attested, although they are less frequent than the non-cognate translations.

3. Analysis of the Slovak data

In the presentation of the analysis, we will start with Slovak. The reason for this will become clear later.

Of the 200 English nouns in *-ation* in our sample, 67 have a cognate Slovak noun in *-ácia* of the type in (3a) that is recorded in KSSJ (2003). This may seem a poor result, but we have to consider that the 200 English nouns are a randomized sample over the nouns in *-ácia* recorded in OED (2000–2022). The only criterion is the recording of a corresponding verb. This means that there are many rare and obsolete nouns in the sample. In fact, there are Slovak cognates such as *lignifikácia* ('lignification') corresponding to nouns that have no occurrences in COCA (2008–2019). In addition, 45 predicted cognates could be found in SNK or by a search on the web. For other cases, Slovak only has non-cognate equivalents, e.g. *prevaha* ('predomination').

As a next step, we considered the verbs in *-ovať*, as in (3b). Table 1 gives the correlation between the recording of the cognate noun and of the corresponding verb in *ovať*.

	Noun in *-ácia* in KSSJ (2003)	**Noun in *-ácia* in corpus**	**No noun in *-ácia* recorded**
Cognate verb in KSSJ (2003)	***41***	*2*	*1*
Cognate verb in corpus	18	23	0
No cognate verb	8	20	87
Total	67	45	88

Table 1: Correlations between Slovak nouns in *-ácia* and corresponding verbs in *-ovať*

In (3a–b), we saw an example where both the noun in *-ácia* and the corresponding verb are recorded in KSSJ (2003). An example where the noun is recorded and the verb only found in the corpus is *inicializácia* ('initialization') and *inicializovať*. For *flagelácia* ('flagellation'), no cognate verb could be found. Table 1 shows that the verb in *-ovať* is strongly dependent on the noun in *-ácia*. Lexicographers also tend to record both the noun and the verb or only the noun rather than only the verb. If the noun in *-ácia* is used but not recorded in the dictionary, this makes the appearance of the verb in *-ovať* more likely, but it is generally not in the dictionary.

Let us now turn to the nouns in *-nie*, as in (3c). The status of these nouns is determined by their largely predictable form and meaning. The relationship between a noun such as (3c) and its corresponding verb in (3b) can be compared to, for instance, the formation of adjectives in *-able* or nouns in *-ing*. In the same way as many English dictionaries will not record them or only give them as run-on entries, dictionaries of Slovak rarely record nouns in *-nie*. For our sample, KSSJ (2003) only contains entries for 5 cognate nouns in *-nie*. For all of these, both the verb in *-ovať* and the noun in *-ácia* are in the dictionary as well. The dependence on the verb is further illustrated by the fact that for all nouns in *-nie* attested in the corpus or on the web also the verb is attested. On this basis, we can safely conclude that the noun in *-nie* occurs only after the formation of the verb.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

These conclusions are in accordance with earlier research which did not use corpus data. Furdík (1978, p. 112) proposes that nouns in *-ácia* were borrowed first. Verbs that were originally derived from these nouns were analysed as motivating the nouns when Slovak speakers became sensitive to the systematic nature of the relation. Furdík calls this *remotivation*. Also Buzássyová (1983, p. 270) observes that the verbs in *-ovať* and the nouns in *-nie* were formed later. They call the verbs *backformations*.

4. Analysis of the German data

As explained in section 1, the situation in German is more complex than in Slovak, because of the different status of the competing suffix *-ung*. Whereas in Slovak, nominalization with *-nie* is highly regular, in a way comparable to the formation of participles, German *-ung* has a degree of irregularity that is more typical of word formation rules. The correlations between the cognates recorded in dictionaries and found in DeReKo are represented in Figure 1.

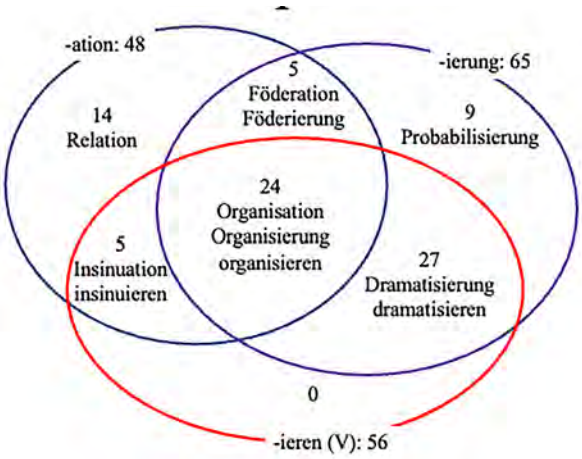


Fig. 1: Correlations between nouns in *-ation*, nouns in *-ierung* and verbs in *-ieren* in German

In Figure 1, the nouns in *-ation* are represented by a blue oval intersecting with the purple oval for the nouns in *-ierung* and the red oval for the verbs in *-ieren*. The numbers in Figure 1 refer to the number of cognates found for the English set of 200 nouns in *-ation* that served as our starting point. There are, for instance, 27 English nouns for which a cognate noun in *-ierung* and a corresponding verb in *-ieren* was found, but no noun in *-ation*. For each class, an example is given. The totals for each of the three sets are indicated as well. That there are no verbs without a nominalization is a consequence of our taking nouns as a starting point.

A first point to be made is that in German there are considerably fewer English nouns for which any cognates were found at all. This may not be immediately obvious from Figure 1, but 116 out of 200 English nouns are not in the figure at all. This compares with 87 for Slovak in Table 1. This number of 116 includes 5 cases where German uses a non-cognate Latinate stem. An example is *anatomization*, for which German has *Präparation* with *präparieren* as a verb and *Präparierung* as an alternative nominalization. In other cases, no Latinate stem is used, e.g. *Vorverurteilung* ('precondemnation').

The most interesting aspect of Figure 1 is what it shows about the competition between *-ation* and *-ierung*. There are about twice as many cognate nouns in *-ierung* without a cor-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

responding noun in *-ation* as there are nouns in *-ation* with a corresponding noun in *-ierung* (36 against 19). The number of cases where both were found is in between (29). It is striking that for the cases with only a noun in *-ation*, fewer than a third have a corresponding verb (5 out of 19), whereas for the cases with only a noun in *-ierung*, three quarters have such a verb (27 out of 36). This suggests that nouns in *-ierung* have a stronger connection to the verb than nouns in *-ation*.

A complex notion that is presupposed in the data represented in Figure 1 is that of a word being *established* in the language. As argued by ten Hacken (2020), there is no empirical object corresponding to named languages. We use *German* to refer to the language of a speech community, but its cognitive basis can only be found in individual speakers. Moreover, the mental lexicon of an individual speaker is highly structured, with some items much more prominent than others. This means that the property *established in German* is gradual in two respects. On one hand, we could try to find out how many speakers classified as members of the German speech community know a particular word. On the other hand, we could try to find out what the position of the word in each speaker's mental lexicon is. Setting thresholds on these two measures is necessarily arbitrary at least to some extent. At the same time, measuring these two parameters is a complex task. Therefore, using frequency in a corpus as an approximation is a convenient patch.

Frequency in a corpus is related both to the proportion of speakers who know a word and to the prominence this word has for them. However, the relationship is indirect. A corpus collects performance data for many speakers, but for some speakers, the chances for their utterances or writings to become part of a corpus are considerably greater than for others. The prominence of a word is reflected to some extent in its frequency of use, but it is difficult to express the relationship between prominence and frequency in a non-circular way.

With these caveats, we analysed the relative frequency of the corresponding words as represented in Figure 1. We first considered the relation between the verbs in *-ieren* and the corresponding nouns in *-ation*. In this analysis we included cases such as *Präparation* ('anatomization') mentioned earlier, so that there are 32 pairs. Of these, there are 2 for which neither the verb nor the noun occurs in the W corpus of DeReKo. For the remaining pairs, in 10 cases the noun in *-ation* is more frequent and in 20 cases the verb in *-ieren*. An example of a case where the noun is more frequent is *Zivilisation* ('civilization'), which is 2.6 times more frequent than *zivilisieren* ('civilize'). An example of a case where the verb is more frequent is *degenerieren* ('degenerate'), which is 3.6 times more frequent than *Degeneration* ('degeneration').

Then we turn to the relation between the verbs in *-ieren* and the corresponding nouns in *-ierung*. Again including cases such as *Präparierung* ('anatomization'), we have 55 pairs here. Of these, 9 pairs have no occurrences for either the verb or the noun in the W corpus of DeReKo. Of the remaining pairs, the noun in *-ierung* is more frequent in 12 cases and the verb in *-ieren* is more frequent in 34 cases. An example of a case where the noun is more frequent is *Dezentralisierung* ('decentralization'), which is 2.3 times as frequent as *dezentralisieren* ('decentralize'). An example of a case where the verb is more frequent is *fixieren* ('fixate'), which is 7.7 times as frequent as *Fixierung* ('fixation'). We see then, that both for nouns in *-ation* and for those in *-ierung*, the verb tends to be more frequent than the corresponding noun.

A legitimate question is of course to what extent the higher frequency of the verb is caused by the fact that the frequency for the nominalizations is divided over two forms. Therefore,

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

we also compared the frequency of the verb in *-ieren* in relation to the combined frequency of the nouns in *-ation* and *-ierung*. There are 60 relevant sets to be compared, of which 7 have no occurrences of any noun or verb in the W corpus of DeReKo. In 22 cases the combined frequency of the nouns is higher than that of the corresponding verb, whereas in 31 cases the verb is more frequent. These figures suggests that there is indeed a less strong predominance of the verb than expected on the basis of the figures for the individual nominalizing suffixes.

In order to interpret the figures for the combination of the two nominalization processes and relate them to the figures of the individual processes, it is worth having a closer look at the competition between *-ation* and *-ierung*. There are 33 cases where the same stem is used for a noun in *-ation* and a noun in *-ierung*. In 2 cases, neither noun occurs in the W corpus of DeReKo. In 5 cases, the noun in *-ierung* is more frequent. An example is *Elektrifizierung* ('electrification'), which has 9,845 occurrences as against 57 for *Elektrifikation*. In 26 cases, the noun in *-ation* is more frequent. An example is (2a, c), where *Organisation* has 801,422 occurrences and *Organisierung* 1,300. The difference in frequency is typically high. The noun in *-ierung* does not occur in the corpus at all in 6 cases and the frequency is between 1 and 10 in 10 cases. A frequency of 10 corresponds to 0.001 per million words.

Returning to the question of whether the split between *-ation* and *-ierung* affects which of the noun or the verb is more frequent, we can be confident to give a negative answer for our sample. The closest call is the case in (4).

(4) a. Isolation 53,744
b. Isolierung 20,340
c. isolieren 115,983

The two nouns in (4a–b) both mean *isolation*. The range of their meanings is not quite identical, but there is a large overlap. They are both established, although the noun in *-ation* is clearly more frequent. The verb in (4c) is, however, much more frequent than the two nouns combined.

The fact that the combination of nouns in *-ation* and *-ierung* yields a less striking contrast between nouns and verbs requires a different explanation. We can summarize the figures as in Table 2.

	Noun more frequent	**Verb more frequent**	**No occurrences**
Noun in *-ation* vs. Verb	***10***	*20*	*2*
Noun in *-ierung* vs. Verb	12	34	9
Sum of nouns vs. Verb	22	31	7

Table 2: Overview of frequency comparisons of nouns in *-ation* and *-ierung* with corresponding verbs

The competition between *-ation* and *-ierung* means that usually one of the two establishes itself, while the other one remains marginal. In Table 2, this is reflected by the fact that in the first column, i.e. the cases where the noun is more frequent than the verb, the number for the combination of the nouns is the sum of the numbers for the individual nouns being more frequent than the verb. This contrasts with the second column. A good example is the triple in (5).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

(5) a. Zivilisation 57,668
b. Zivilisierung 1,353
c. zivilisieren 21,807

As indicated by the frequencies in (5), the established noun corresponding to *civilization* is (5a). The alternative noun (5b) is used only in the process reading. It only has a chance because in some contexts, it is desirable to foreground this reading. The noun (5a) is more than twice as frequent as the verb (5c). The noun (5b) is much less frequent than the verb. What is reflected in the column with more frequent verbs in Table 2 is then that the comparison between (5b) and (5c) disappears as an item to be counted when we combine the values for the two nouns in (5a) and (5b).

Another point to be considered here is the use of non-cognate nouns instead of the nouns in *-ation* or *-ierung*. An example is the triple for *petrification*, given with its frequencies in (6).

(6) a. Petrifikation 5
b. Petrifizierung 26
c. Versteinerung 3,466

Although both forms (6a–b) are attested, they are definitely rare and much less common than (6c). The problem with counting such cases as (6) is that it is often not easy to establish that the native form has exactly the same lexical meaning. In (6) this is relatively obvious, but (7) is less straightforward.

(7) a. Annotation 151
b. Annotierung 4
c. Anmerkung 58,088

Both nouns (7a–b) are attested as equivalents to *annotation*. The difference between them is similar to the one in (5a–b), but they are much rarer. The question is to what extent the use of the native equivalent (7c) is responsible for the low frequency of (7a–b). Collins (1999) gives "note" and "remark, comment" as translations of *Anmerkung*, but "Anmerkung" as a translation of *annotation*.

With these caveats in place, it is still worth considering to what extent alternative words are used for nouns in *-ation*. In fact, only 10 of the 48 nouns in *-ation* do not have a synonym without the cognate stem and in 33 cases, this synonym is more frequent. The frequencies in the examples (6–7) are not at all untypical.

The German data suggest that nouns in *-ation* are relatively marked. For the 200 English nouns in our sample, only 48 have a cognate noun in *-ation* and in most cases, a more frequently used synonym without the cognate stem exists. Nouns in *-ierung* are competing with corresponding nouns in *-ation*. The nouns in *-ierung* often prevail in frequency and are more often in correspondence with a verb in *-ieren*.

5. Frequency data for Slovak

In the discussion of the Slovak data in section 3, we only considered the attestation of cognate verbs and nouns in KSSJ (2003) and SNK. As we used frequency data in our analysis of the corresponding German cases, it is legitimate to ask to what extent taking into account frequency data for Slovak modifies the conclusions we reached earlier. Panocová (2017) presents an earlier study of the frequency of nouns in *-ácia* and their corresponding verbs

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

in *-ovať* and nouns in *-nie*. This study was based on nouns in *-ácia* selected from the frequency lists of SNK.

We first considered the relation between the verbs in *-ovať* with a Latinate stem and the corresponding nouns in *-ácia*. For Slovak, we not only looked for these in the dictionary and in the corpus, but also by means of a general web search. The web search was not used to establish the frequency of a word, but if the predicted word occurs in a Slovak text, it is taken as an attested word. There are 90 relevant verbs, of which 34 have no occurrences in SNK. The verbs that do not occur in SNK are sometimes recorded in KSSJ (2003), e.g. *inundovať* ('inundate'). For 85 verbs, we found a corresponding noun in *-ácia* (cf. Table 1). In 36 cases, the noun is more frequent than the verb. In addition, in 25 cases, only the noun is attested in SNK. In 18 cases, the verb is more frequent. In the remaining 6 cases, neither the verb nor the noun occurs in SNK. These data are in line with the hypothesis that the verbs in *-ovať* are often backformations, i.e. they depend on the noun in *-ácia*.

Then we turned to the relation between the verbs in *-ovať* and the corresponding nouns in *-nie*. We found 60 pairs of a Latinate verb in *-ovať* with a corresponding noun in *-nie*. This is a clearly lower number than the 85 pairs with *-ácia*. There are only 2 nouns with a higher frequency in SNK than the corresponding verb, *falšovanie* ('falsification') and *aprobovanie* ('approbation'). Interestingly, neither of these has an entry in KSSJ (2003), although the noun in *-ácia* and the verb in *-ovať* do. This is particularly remarkable in the case of (8).

(8) a. falzifikácia 360
b. falšovanie 7,223
c. falšovať 2,165

Also in the case of (8), the frequency data are in line with the conclusion we drew in section 3 that nouns in *-nie* depend on the corresponding verb in *-ovať*.

Finally, we considered the cases where nouns of both types, *-ácia* and *-nie*, were in competition. There are 56 such pairs in our sample. For 49 of these, the noun in *-ácia* is more frequent. In 3 cases, neither noun is attested in SNK. Only 3 nouns in *-nie* are more frequent than the corresponding noun in *-ácia*. One of them is (8b). Another case is given in (9).

(9) a. transportácia 1
b. transportovanie 60
c. transportovať 2,888
d. transport 14,865

Although the noun in *-nie* in (9b) is more frequent than the one in *-ácia* in (9a), the established noun is (9d). Interestingly, KSSJ (2003) has entries for (9a, c–d), but not for (9b). It is also interesting to note that of the 56 nouns in *-nie* for which a counterpart in *-ácia* is attested in our sample, 34 have a frequency of at most 7 in SNK and only 6 a frequency over 160, which corresponds to 0.1 per million.

On this basis, it seems safe to conclude that nouns in *-ácia* have a very strong position in the Slovak lexicon. They generally seem to underlie the corresponding verb and have very little competition from other nouns. Nouns in *-nie* are implied by the verb and used in a way not unlike the nominalized infinitive in German. They are generally much less frequent, but they can be used to highlight the process reading, as in (3). There are only 2 cases in our sample where a different noun is used as a competitor for the noun in *-ácia*. One is given in (9d), the other is *detox*, a less frequent variant of *detoxifikácia* ('detoxification'). In both cases

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

we have a shortened form that is probably a loan from English. This suggests that once a noun in *-ácia* is in the lexicon, it excludes the formation of competing nouns.

6. Comparison of German and Slovak

When we compare the constellation of *-ation* in German and *-ácia* in Slovak, the first striking difference is the status of the competing suffixes, *-ung* in German and *-nie* in Slovak. In both languages, these suffixes have a wider distribution than the corresponding Latinate suffixes, i. e. they can apply to a larger set of verbs. However, *-ung* and *-nie* are different in a property that is often labelled *productivity*. We can say that *-ung* is less productive than *-nie*. On closer inspection, there is also another difference in productivity, one that directly affects *-ation* and *-ácia*.

Productivity of a morphological rule is an intuitively clear concept, which is, however, difficult to pin down exactly. Bauer (2001) gives an overview of the discussion. Here we will adopt Corbin's (1987, p. 176–178) analysis, which distinguishes three separate aspects, *régularité*, *disponibilité* and *rentabilité*. The first of these, *régularité*, refers to the extent that the form and meaning of the resulting word is predictable. The lack of this property is an important criterion for the inclusion of words in a dictionary. It is in this sense that Slovak *-nie* is more productive than German *-ung*. Whereas the Slovak noun in *-nie* can always be used and always has a process reading, the German noun in *-ung* is not always available and may also have other than process readings. The second type, *disponibilité*, answers the question whether a particular word formation process can be used to produce new words. A positive answer is a condition for using the process in an etymological account. Obviously, the availability of a rule may change over time and is dependent on individual speakers. The third type, *rentabilité*, refers to the quantitative aspect of productivity. It is the productivity-related information one can retrieve from a corpus. However, as observed by Baayen (1992), it is not the number of tokens for a particular word, but rather the number of types formed by a morphological process that should be taken as a basis, because a new type indicates the active use of the rule. Clearly, these three types are related to each other, but it is useful to maintain the distinction, because an increase in one type of productivity does not automatically result in an increase in the other types.

In Slovak, nouns in *-ácia* were generally borrowed. The data we collected support the traditional view formulated by Furdík (1978) and Buzássyová (1983) that the verbs in *-ovať* are based on the reanalysis of the borrowed nouns in *-ácia* as complex. For verbs in *-ovať*, a corresponding noun in *-nie* can be formed. This constellation is reflected in the attestation of nouns and verbs. The noun in *-ácia* is primary, the verb in *-ovať* depends on this noun and the noun in *-nie* depends on the verb. When a formation depends on another word, it is available for use whenever the need arises.

In German, nouns in *-ation* were also generally borrowed. However, their relation to the corresponding verbs in *-ieren* is not the same as in Slovak. The frequent occurrence of a pair of cognate verb in *-ieren* and noun in *-ierung* suggests that in many cases the verb in *-ieren* was borrowed independently. This suggestion is reinforced by the observation that in the majority of cases, the verb is more frequent than the corresponding noun. This means that in German, when a noun in *-ation* and a corresponding verb in *-ieren* both exist, they are connected in the mental lexicon, but it is not the case that the verb is formed automatically

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

when the noun exists. This also explains that *-ung* is much more prominent in German than *-nie* is in Slovak.

In conclusion, we can say that the etymology of internationalisms in *-ation* and *-ácia* is different in German and in Slovak. In Slovak, the data are compatible with the analysis that nouns in *-ácia* are generally borrowings and corresponding verbs are backformations, in German we have to assume that also many of the verbs in *-ieren* were borrowed. Therefore, they are etymologically not backformations based on the noun in *-ation*. The nouns in *-ation* may be analysed as derived from the verbs or as separate borrowings.

In formulating an etymology for a dictionary, it is often difficult to assess whether the word in question was borrowed or resulted from word formation. Therefore, it is important to have such general hypotheses in mind when deciding how to present its origin. Here we formulated some data-supported hypotheses that can be used as default assumptions in etymologies.

References

Baayen, H. (1992): Quantitative aspects of morphological productivity. In: Booij, G./van Marle, J. (eds.): Yearbook of morphology 1991. Dordrecht, pp. 109–149.

Bauer, L. (2001): Morphological productivity. Cambridge.

Braun, P. (1990): Internationalismen – gleiche Wortschätze in europäischen Sprachen. In: Braun, P./Schaeder, B./Volmert, J. (eds.): Internationalismen: Studien zur interlingualen Lexikologie und Lexikographie. Tübingen, pp. 13–33.

Buzássyová, K. (1983): Konkurencia slovotvorných typov s formantmi -(*iz*)*ácia*, -(*ova*)*nie*. In: Slovenská reč 48 (5), pp. 268–277.

Buzássyová, K. (1991): Opakovaná internacionalizácia a problém identifikácie morfologických a lexikálnych jednotiek. In: Jazykovedný časopis 42 (2), pp. 89–104.

COCA (2008–2019): The Corpus of Contemporary American English. Edited by Mark Davies. http://corpus.byu.edu/coca/ (last access: 09-05-2022).

Collins (1999): Collins German-English English-German dictionary. Unabridged. 4th edition. Glasgow.

Corbin, D. (1987): Morphologie dérivationnelle et structuration du lexique. Tübingen.

DeReKo (2005–2022): Deutsches Referenzkorpus, Mannheim: Leibniz-Institut für Deutsche Sprache. www.ids-mannheim.de/DeReKo (last access: 09-05-2022).

Duden (2011–2022): Duden online. Berlin: Bibliographisches Institut. https://www.duden.de (last access: 09-05-2022).

Fleischer, W./Barz, I. (2012): Wortbildung der deutschen Gegenwartssprache. 4. Auflage. Berlin.

Furdík, J. (1978): Slovotvorná motivovanosť slovnej zásoby v slovenčine. In: Mistrík, J. (ed.): Studia Academica Slovaca 7: Prednášky XIV. letného seminára slovenského jazyka a kultúry. Bratislava, pp. 103–115.

ten Hacken, P. (2020): Norms, new words, and empirical reality. In: International Journal of Lexicography 33 (2), pp. 135–147.

ten Hacken, P./Panocová, R. (2022): The suffix *-ation* in English. In: Arbeiten aus Anglistik und Amerikanistik 47 (1), pp. 29–57.

Horecký, J. et al. (1989): Dynamika slovnej zásoby súčasnej slovenčiny. Bratislava.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

KSSJ (2003): Krátky slovník slovenského jazyka. 4th edition. Edited by Kačala, J./Pisárčiková, M./Považaj, M. Bratislava.

OED (2000–2022): Oxford English dictionary. 3rd edition. Edited by John Simpson and Michael Proffitt. Oxford. www.oed.com (last access: 09-05-2022).

Panocová, R. (2017): Internationalisms with the suffix *-ácia* and their adaptation in Slovak. In: Litta, E./Passarotti, M. (eds.): Proceedings of the Workshop on Resources and Tools for Derivational Morphology (DeriMo), 5–6 October 2017, Milano, Italy. Milano, pp. 61–72.

SNK (2020): Slovenský národný korpus. prim-9.0-public-sane. Bratislava: Jazykovedný ústav Ľ. Štúra SAV. bonito.korpus.sk (last access: 09-05-2022).

Waszakowa, K. (2003): Internacionalizacija: Zapadnoslavianskije jazyki. Przejawy tendencji do internacjonalizaji w systemach słowotwótwórczych języków zachodnosłowiańskich. In: Ohnheiser, I. (ed.): Komparacja systemów i funkcjonowanie współczesnych języków słowiańskich. 1. Słowotwórstwo/Nominacja. Innsbruck/Opole, pp. 78–102.

Contact information

Pius ten Hacken
Leopold-Franzens-Universität Innsbruck
pius.ten-hacken@uibk.ac.at

Renáta Panocová
Pavol Jozef Šafárik University in Košice
renata.panocova@upjs.sk

Acknowledgements

This work was supported by the Scientific Grant Agency of the Ministry of Education, Science, Research and Sport of the Slovak Republic and Slovak Academy of Sciences VEGA under the project No VEGA 1/0130/21. We would like to thank Lisa-Marie Lang for her work on German data collection and Lukáš Lukačín for his work on the Slovak data.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Neologisms and Lexicography

XX EURALEX INTERNATIONAL CONGRESS
DICTIONARIES AND SOCIETY
12-16 July 2022, Mannheim, Germany

Ieda Maria Alves/Bruno Maroneze

FROM SOCIETY TO NEOLOGY AND LEXICOGRAPHY

Relationships between morphology and dictionaries

Abstract This paper aims at verifying if the most important online Brazilian Portuguese dictionaries include some of the neologisms identified in texts published in the 1990s to 2000s, formed with the elements *ciber-*, *e-*, *bio-*, *eco-* and *narco,* which we refer to as *fractomorphemes* / *fracto-morphèmes*. Three online dictionaries were analyzed (*Aulete*, *Houaiss* and *Michaelis*), as well as *Vocabulário Ortográfico da Língua Portuguesa* (*VOLP*). We were able to conclude that all three dictionaries and VOLP include neologisms with these elements; Michaelis and VOLP do not include separate entries for bound morphemes, whereas Houaiss includes entries for all of them and Aulete includes entries for *bio-*, *eco-* and *narco-*. Aulete also describes the neological meaning of *eco-* and *narco-*, whereas Houaiss does not.

Keywords Fracto-morphèmes; neologisms in Brazilian Portuguese; Brazilian Portuguese dictionaries

1. Introduction

The aim of this study is to verify if Brazilian Portuguese dictionaries, which are currently available online, incorporate neologisms identified in printed press texts published in the 1990s and 2000s.

Our analysis takes into account some morphemes, observed since the 1990s, that have been reflecting new tendencies of contemporary global (and, in particular, Brazilian) society – *ciber-*, *e-* (Information Technology, Computing); *bio-*, *eco-* (healthy, sustainable life and society) – and some of its problems (*narco-*) (cf. Alves, 2004, 2007, 2009).

These morphemes have been termed differently. Sometimes they are termed *prefixes*, because they occupy the first position in the word, but they have also received specific denominations. In English, they are referred to as "splinters", a term thus defined by Bauer (2004, pp. 95 f.): "A splinter is a fraction of a word, arising in a blend, then used as an affix to create more words, as *-nomics* in *Reaganomics*, *Clintonomics*, *Thatchernomics* and so on."

According to Tournier (1985, p. 86), this term was first employed by Adams (1973, pp. 188-98), in order to distinguish among affixes, compound elements and amalgamated elements. This author mentions, as an example, the word *microscope*, in which *-scope* is a splinter. In Brazilian linguistics, authors such as Gonçalves (2016) also use this English term.

Concerning the French language, these elements were studied by Tournier (1985) as "éléments affixés à gauche" (left-hand affix elements), such as *eco-*, *macro-*, *micro-*, *multi-*, *nano-*.

About *eco-* (which we study in this paper) the author writes, in a footnote: "ce fracto-morphème connaît une productivité croissante depuis la fin des années 1960" (this fractomorpheme has experienced increasing productivity since the late 1960s) (p. 94). The term *fractomorphème*, widely used in French and also in European Portuguese as the calque *fractomorfema*, is Tournier's coinage to translate the English term *splinter*. "Un fracto-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

morphème est donc un fragment de lexie qui la représente dans un mot construit" (A fractomorpheme is, therefore, a fragment of a word that represents it in a constructed word) (Tournier 1985, p. 86).

Jean-François Sablayrolles, in his studies on neology in French, has employed the term *fractolexème*, considering it more suitable than *fractomorphème*. Through acknowledging that *fractolexème* represents a kind of compound lexical unit, he includes *fractocomposition* (*fractocompounding*) in his typology of neologisms. According to his definition, *fractocomposition* "ne se distingue de la composition 'normale' que par le fait qu'un des éléments constitutifs n'est pas un mot complet, mais un fragment de celui-ci, qui vaut, sémantiquement, pour l'ensemble" ([fractocompounding] distinguishes itself from 'normal' compounding only because one of its elements is not a complete word, but a fragment of it, that has the semantic value of the whole) (2017, p. 54).

In this text, we have chosen to term these elements *fractomorphemes*, following the French tradition as well as some Portuguese and Brazilian authors, such as Lino (1990) and Bizzocchi (2021).

1.1 Description of the morphemes

The fractomorphemes here analysed represent elements which have been widely employed since the beginning of the 1990s. *Ciber-*, *e-*, *bio-*, *eco-* and *narco-* picture different aspects which have been posed as representative of the society of this period and of the following decades.

From the technological point of view, there is a development of already used technologies such as the Internet as weel as the intensification of the use of cell phones, which have become smaller. The possibilities of virtual communication are considerably expanded, which is expressed by the fractomorphemes *ciber-* (from Eng. *cybernetics*, according to Houaiss) and *e-* (< Eng. reduction of *electronic*, according to Houaiss).

Regarding the environment, the 1990s stood out in relation to the previous periods, due to the importance assumed by the Earth Summit, a conference organized by the UN in 1992, in Rio de Janeiro, also known as Eco 92, which was attended by representatives of several countries. The Eco 92 contributed to bringing together the concepts of society, economy and sustainability, seeking to raise awareness about the importance of taking care of nature so that future generations can enjoy its resources (Takeda 2009). This concern for the care of the environment was reflected in new usages of the elements *bio-* (< Greek *bíos*, 'life', according to Houaiss) and *eco-* (< Greek *oikos*, 'house, household, family', according to Houaiss).

The 1990s also saw a change in the relationship between drug trafficking and the policies of different nations regarding the circulation and use of drugs in their territories. As Vilela (2015) points out, "Brazil's stance on the issue of drugs changed significantly in the 1990s, when a series of legislative and institutional apparatuses were established with the aim of fighting drug trafficking, based on the identification of threats associated with this crime" ("A postura do Brasil com relação ao tema das drogas mudou significantemente nos anos 1990, quando uma série de aparatos legislativos e institucionais foram estabelecidos com o objetivo de combater o tráfico de drogas, a partir da identificação das ameaças associadas a este crime"). In this context, the fractomorpheme *narco-*, designating *narcotics*, began to be used in various formations linked to drug trafficking.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

It is important to point out that, although the elements *e-* and *ciber-* may be considered 'pure' fractomorphemes, the elements *bio-*, *eco-* and *narco-* also have a 'classical' usage as compound elements. However, there is a significant difference in meaning between the 'classical' compound elements and their respective fractomorphemes. In the case of *bio-*, it is present in compounds such as *biologia* 'biology', *biografia* 'biography', in which it means 'life'. As a fractomorpheme, instead of meaning 'life' or 'life-related', it means 'related to biology' or 'biological', in words such as *biodiversidade* 'biodiversity, biological diversity' and *biopirataria* 'biopiracy, biological piracy'.

In the case of *eco-*, it forms *ecologia* 'ecology' and *economia* 'economy', where it means 'house, household'; as a fractomorpheme, it has the neological meaning 'ecological', 'environment-friendly' in formations such as *ecoturismo* 'ecological tourism', *ecoproduto* 'environment-friendly product'.

The compound element *narco-* has the meaning 'numbness, torpor', in compounds such as *narcótico* 'substance that causes numbness, narcotic', *narcolepsia* 'narcolepsy'. As a fractomorpheme, it means 'drug-related', such as *narcotráfico* 'drug trafficking' and *narcoeconomia* 'the economy of drug dealing'.

In regards to the etymology of these fractomorphemes, the morpheme *e-* is clearly a loan from English; the other four may be analysed as reductions or truncations from larger forms (*cibernética* 'cybernetics' for *ciber-*, *biológico* 'biological' for *bio-*, *ecológico* 'ecological' for *eco-* and *narcótico* 'narcotic' for *narco-*. But it is certain that their usage in Portuguese is heavily influenced by their usage in other languages (especially English, Spanish and French); thus, all of them may be considered loans to some extent.

2. Methodology

Our corpus of neologisms is based on the results of Project TermNeo (*Observatório de neologismos do português brasileiro contemporâneo* – Observatory of contemporary Brazilian Portuguese neologisms). This is a project with the scope of collecting and investigating neologisms present in contemporary Brazilian press (newspapers and magazines) since January 1993 (Alves, 2012; Alves/Maroneze 2021). This project was proposed, similarly to other neology observatories in the Romance languages, following the model of the Laboratory of Lexicological Analysis at the Center for Applied Linguistics (*Laboratoire d'Analyse Lexicologique du Centre d'Etudes de Linguistique Appliquée*) of the Université de Franche-Comté (Besançon, Franche-Comté, France), conceived by the lexicologist and lexicographer Bernard Quemada, in the early 1960s, for the detection of neologisms used in the French language press and their insertion in French language dictionaries.

Project TermNeo, as most projects that describe Romance language neologisms extracted from journalistic language, has considered the absence of a word in a set of contemporary lexicographic works and, more recently, other materials (an exclusion corpus) as the most important criterion for a lexical unit to be classified as a neologism. More recently, the development of Information Technology and Corpus Linguistics has allowed us to improve this methodology using automated techniques.

The data collected by Project TermNeo have been extracted from the highest-circulation Brazilian newspapers (*Folha de S. Paulo*, *O Globo*) and magazines (*Veja*, *IstoÉ*, *Época*), since 1993. The typology adopted by Project TermNeo partially follows the typology initially established by Guilbert (1975), plus some contributions observed in the typologies of Cabré

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

(2006, 2010) and Sablayrolles (2000, 2017), plus our own reflections: vernacular morphosyntactic neologisms, which include derivation and compounding processes (prefixed words, suffixed words, coordinative compounds, subordinative compounds and syntactic or syntagmatic compounds); semantic neologisms; other processes such as formations with acronyms, truncations, splinters and blendings; and loanwords.

This systematic observation of neologisms has shown that the presence of a neologism not only implies the addition of a new lexical unit in the lexicon of the language but can also point out the existence of a new semantic or morphosyntactic function related to this new word or one of its formative morphemes (as it was studied in Alves/Maroneze 2021).

Through the systematic observation of neologisms from 1993 onwards, we were able to identify the development of a recent pattern of word-formation in Brazilian Portuguese: words formed with elements such as *ciber-*, *e-*, *bio-*, *eco-* and *narco-*, which share some traits with traditional prefixes, but nowadays are best described as fractomorphemes, as already mentioned in the introduction of this paper.

In order to verify if contemporary Brazilian Portuguese dictionaries describe this new pattern, we chose three online dictionaries (described in the following section). The following questions guided our research:

a) Do the dictionaries include separate entries for these fractomorphemes? If so, do they mention their neological meanings?
b) Do the dictionaries include neologisms formed with these fractomorphemes? If so, how do they describe their structure?

2.1 Description of the dictionaries

The three lexicographical works analyzed here, *Houaiss*, *Aulete* and *Michaelis*, are among the most important Brazilian Portuguese general dictionaries. The main reasons for choosing them were: (i) they are among the largest dictionaries for Brazilian Portuguese which present more than 150,000 entries; (ii) all three have online versions; (iii) their latest versions include some neologisms (as it is shown in this study).

Aulete Dictionary was named after Francisco Júlio de Caldas Aulete, the conceiver of the work. Its first edition was published in Lisbon in 1881; its first Brazilian version was published in Rio in 1958. The current version, which was consulted for this study, is the online version. According to Aulete website (http://www.aulete.com.br/site.php?mdl=aulete_digital&op=o_que_e), the dictionary has more than 818,000 entries, meanings and idioms (*'Mais de 818 mil verbetes, definições e locuções'*).

Houaiss Dictionary was named after Antônio Houaiss, the creator and founder of the work. Its first printed edition came in 2001 and is no longer in print. A smaller version was published in 2009 and is sold in print and electronic (CD-ROM) formats. The foreword (Houaiss; Villar 2009, p. XI) informs that its nomenclature contains about 146,000 entries (from the 230,000 in the 2001 full version). There is also the online version, available in online format for subscribers of *Universo OnLine* (UOL) (https://houaiss.uol.com.br/corporativo/apps/uol_www/v5-4/html/index.php#1). The website does not mention the total number of entries.

Michaelis Dictionary is named after Henriette Michaelis, a German lexicographer from the 19th century who published dictionaries together with her sister, the famous philologist Carolina Michaelis de Vasconcelos. In the 1950s, Melhoramentos publishing house acquired

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

the rights of the *Michaelis* brand and, since then, has published many bilingual and monolingual dictionaries, most of which are freely available online (https://michaelis.uol.com.br/).

The printed version of the Michaelis Portuguese Dictionary was published in 1999. According to the website (https://michaelis.uol.com.br/moderno-portugues/), the revised version of this work was concluded in 2015 and it is available only in digital format, containing around 167,000 entries.

The dictionaries have certain traits in common: all three have printed and online versions and all three have mini versions for students, made especially for the *Programa Nacional do Livro e do Material Didático* (PNLD – a Brazilian governmental program for school books – https://www.fnde.gov.br/programas/programas-do-livro). They differ in the fact that Aulete has a much older publishing tradition and, in its online version, it also encourages users' participation as collaborators for its development.

In addition to these three dictionaries, we also use data provided by VOLP (*Vocabulário Ortográfico da Língua Portuguesa* – Orthographic Vocabulary of the Portuguese Language), in its 6th edition, since 2021 (also available online – https://www.academia.org.br/nossa-lingua/busca-no-vocabulario), which, despite being only an orthographic dictionary, brings 382,000 entries, listing their respective word class labels and other additional information.

3. Analysis of the data

The data from Project TermNeo demonstrate some formations with these elements that are sometimes incorporated into the dictionaries. Here we describe both the data from the project and the dictionary entries found for each of the fractomorphemes.

3.1 *ciber-*

Some of the examples with the element *ciber-* observed at TermNeo's neologism database are: *ciberapresentador*, *cibercafé*, *ciberciúme*, *cibercondríaco*, *cibercrente*, *cibercrime*, *cibercrítico*, *cibercultura*, *ciberdemocracia*, *ciberdetetive*, *ciberempresário*, *ciberespaço*, *ciberguerra*, *ciberguerrilheiro*, *ciberlaranja*, *cibermano*, *cibernauta*, *ciberobra*, *cibertecnologia*, *ciberterapia*.

This morpheme, sometimes spelled *cyber-*, as it is in English, has its own entry only in Houaiss (classified as a compounding element), which mentions many examples (some of them with their own entries): *ciberataque* 'cyberattack', *cibercafé* 'cybercafe', *cibercolapso* 'cybercollapse', *ciberespacial* 'cyberspace' (adjective), *ciberespaço* 'cyberspace' (noun), *ciberpirata* 'cyberpirate', *ciberpirataria* 'cyberpiracy', *ciberterrorismo* 'cyberterrorism', *ciberterrorista* 'cyberterrorist'.

In Michaelis, the following are included as entries: *cibercriminoso* 'cybercriminal', *cibercultura* 'cyberculture', *ciberdependência* 'cyberaddiction, cyberdependency', *ciberdependente* 'cyberaddicted', *ciberespacial* 'cyberspace adj', *cibermundo* 'cyberworld', among others.

Aulete includes, as entries, *ciberespacial* 'cyberspace adj.', *ciberespaço* 'cyberspace n.', *ciberladrão* 'cyberthief', *cibermaníaco* 'cybermaniac', *ciberpornografia* 'cyberpornography', *ciberpropaganda* 'cyberadvertising', among others.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

3.2 *e-*

The morpheme *e-*, also employed in Information Technology and Computing, may be exemplified by the following neologisms (from TermNeo's database): *e-analfabetismo*, *e-book*, *e-cinema*, *e-comércio*, *e-contracheque*, *e-cultura*, *e-curioso*, *e-livro*, *e-trailer*, *e-voto*, among others. It is included as an entry only in Dictionary Houaiss (classified as a compounding element), with the explanation that it is employed in compounds whose second element is usually an English word (*e-book*, *e-mail*, *e-business*) designating activities and products related to the Internet. Aulete includes entries for *e-book* and *e-mail* and Michaelis includes entries for *e-mail*, *e-pub*, *e-tag* and *e-book*.

3.3 *bio-*

Among the formations with *bio-* observed at TermNeo's database, there are neologisms such as *biopirataria*, *bioarqueólogo*, *biocampeão*, *biochip*, *biodiesel*, *biodiversidade*, *bioforma*, *bionauta*, *biopirata*, *bioprospecção*. This morpheme has its own entry in Houaiss and Aulete, but not in Michaelis. Houaiss classifies it as a compounding element, but Aulete says it can be prefixal (as in *biologia* 'biology') or suffixal (as in *micróbio* 'microbe'). Traditionally, it forms words with the meaning 'life', but in neologisms it usually has the meaning of 'healthy' or 'eco-friendly', such as *bioforma* 'bio- + shape, form', *bioorgânico* 'bio-organic'. None of the dictionaries describes the neological meaning of this morpheme. Houaiss includes *bioagricultura* 'environment-friendly agriculture' and *biopirataria* 'biopiracy'; Michaelis also includes *biopirataria*. So far, Aulete only includes the neologism *biodiesel*.

Unlike French, in which the element *bio* may be used as an adjective (*nourriture bio* 'healthy food', *legumes bio* 'eco-friendly vegetables'), in Brazilian Portuguese this usage is unattested.

3.4 *eco-*

The morpheme *eco-* has a meaning close to that of *bio-*, referring to eco-friendly facts. Some of the examples from TermNeo's database are: *ecobesteira*, *ecobife*, *ecodesigner*, *ecofuturista*, *ecoproduto*, *ecorregião*, *ecoturismo*, *ecossonda*, *ecovisitante*. In neologisms formed with this morpheme, sometimes a playful aspect is also observed in words that suggest an exaggeration on environmental concerns: *ecobrigão* from *brigão* 'a person who fights for anything', *ecochato* from '*chato*' as in a 'boring person', *ecoxiita* from *xiita, meaning* 'extremist, fundamentalist'…

It is included as an entry in Houaiss and Aulete. Houaiss describes its meaning as 'house', 'habitat', 'family'; Aulete does the same, but also includes the neological meaning 'ecological', mentioning neologisms such as *ecoturismo* 'ecotourism' and *ecoturista* 'ecotourist'. All three dictionaries include entries for *ecoturismo* and *ecoturista*.

3.5 *narco-*

The morpheme *narco-* is observed, in TermNeo's database, in some formations related to illegal drug trafficking and, specifically, to the relationship between drug trafficking and politics, such as: *narcocorrupção* 'narcocorruption', *narcoeconomia* 'narcoeconomy', *narco-*

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

parlamentar 'narcocongressman', *narcopolítica* 'narcopolitics'. It is included as an entry in Houaiss and Aulete, but not in Michaelis. Both dictionaries describe its meaning as 'numbness', but Aulete includes the neological meaning 'related to narcotrafficking'. All three dictionaries include an entry for *narcotráfico* 'narcotrafficking'; Houaiss and Aulete include also entries for *narcoterrorismo* and *narcoterrorista*; Aulete also includes the neologism *narcotúnel* 'narco-tunnel, tunnel used for drug trafficking'.

We have also identified that *Vocabulário Ortográfico da Língua Portuguesa* (ABL, 2021) lists neologisms formed with all of the analyzed morphemes, such as: *ciberbullying* 'cyber-bullying', *cibercrime* 'cyber-crime'; *e-book*, *e-commerce*, *e-mail*; *ecopolítica* 'eco-politics', *ecoterrorismo* 'eco-terrorism'; *bioconsciência* 'bio-conscience', *bioecologia* 'bio-ecology', *biofertilizante* 'bio-fertilizer'; *narcotraficante* 'drug dealer', *narcotráfico* 'narcotrafficking'.

4. Final remarks

Our analysis has identified that, in general, all four lexicographical works include at least some neologisms formed with these morphemes. The only dictionary that includes entries for all of these morphemes is Houaiss, although it does not mention the neological meaning of *bio-*, *eco-* and *narco-*. Aulete includes most morphemes (with the exception of *ciber-* and *e-*), and it even mentions the neological meanings of *eco-* and *narco-*. Michaelis does not include any entries for these morphemes. The following table summarizes the data.

	Houaiss	**Aulete**	**Michaelis**	**VOLP**
ciber-	- Includes a separate entry for the fractomorpheme and classifies it as a compounding element - Includes many neologisms, such as *ciberataque*, *cibercafé*, *cibercolapso*, *ciberespacial*, *ciberespaço*, *ciberpirata*, *ciberpirataria*, *ciberterrorismo*, *ciberterrorista*	- Does not include a separate entry for the fractomorpheme - Includes some neologisms, such as *ciberespacial*, *ciberespaço*, *ciberladrão*, *cibermaníaco*, *ciberpornografia*, *ciberpropaganda*	- Does not include a separate entry for the fractomorpheme - Includes some neologisms, such as *cibercriminoso*, *cibercultura*, *ciberdependência*, *ciberespacial*, *cibermundo*	- Does not include a separate entry for the fractomorpheme - Includes some neologisms, such as *ciberbullying*, *cibercrime*
e-	- Includes a separate entry for the fractomorpheme and classifies it as a compounding element - Includes many neologisms, such as *e-banking*, *e-book*, *e-business*, *e-ink*, *e-mail*, *e-reader*	- Does not include a separate entry for the fractomorpheme - Includes some neologisms, such as *e-book* and *e-mail*	- Does not include a separate entry for the fractomorpheme - Includes some neologisms, such as *e-book*, *e-mail*, *e-pub*, *e-tag*	- Does not include a separate entry for the fractomorpheme - Includes some neologisms, such as *e-book*, *e-commerce*, *e-mail*

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

	Houaiss	Aulete	Michaelis	VOLP
bio-	- Includes a separate entry for the fracto-morpheme and classifies it as a compounding element - Does not describe its neological meaning - Includes *bioagricultura* and *biopirataria*	- Includes a separate entry for the fracto-morpheme and classifies it as either a prefix (*biologia* 'biology') or a suffix (*micróbio* 'microbe') - Does not describe its neological meaning - Includes the neologism *biodiesel*	- Does not include a separate entry for the fractomorpheme - Includes *biopirataria*	- Does not include a separate entry for the fractomorpheme - Includes some neologisms, such as *bioconsciência*, *bioecologia*
eco-	- Includes a separate entry for the fracto-morpheme and classifies it as a compounding element - Does not describe its neological meaning - Includes *ecoturismo* and *ecoturista*	- Includes a separate entry for the fracto-morpheme and classifies it as a compounding element - Describes its neological meaning 'ecological' - Includes *ecoturismo* and *ecoturista*	- Does not include a separate entry for the fractomorpheme - Includes *ecoturismo* and *ecoturista*	- Does not include a separate entry for the fractomorpheme - Includes *ecoturismo* and *ecoturista*
narco-	- Includes a separate entry for the fracto-morpheme and classifies it as a compounding element - Does not describe its neological meaning - Includes *narcotráfico* 'narcotrafficking', *narcoterrorismo* 'narcoterrorism', *narcoterrorista* 'narcoterrorist'	- Includes a separate entry for the fracto-morpheme and classifies it as a compounding element - Describes its neological meaning 'narcotic' - Includes *narcotráfico* 'narcotrafficking', *narcoterrorismo* 'narcoterrorism', *narcoterrorista* 'narcoterrorist', *narcotúnel* 'narcotunnel'	- Does not include a separate entry for the fractomorpheme - Includes *narcotráfico* 'narcotrafficking'	- Does not include a separate entry for the fractomorpheme - Includes *narcotráfico* 'narcotrafficking'

Table 1: Summary of the data

Now we are able to answer the questions posed at section 2:

a) Do the dictionaries include separate entries for these fractomorphemes? If so, do they mention their neological meanings?

Houaiss includes separate entries for all of them; Aulete does not include separate entries for *e-* and *ciber-* (possibly because they do not correspond to 'classical' compound elements). However, Aulete is the only one that mentions the neological meanings of two of them (*eco-* and *narco-*).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Michaelis and VOLP do not include separate entries for the fractomorphemes, apparently because they include only 'full' words.

b) Do the dictionaries include neologisms formed with these fractomorphemes? If so, how do they describe their structure?

All lexicographical works include at least one neologism for each fractomorpheme.

As already highlighted in Alves/Maroneze (2021), even though the dictionaries' main goal is to describe individual lexical items, it does not prevent them from describing grammatical facts. The works here analyzed, while relatively up-to-date in terms of including neologisms, are still incomplete concerning the description of morphological elements.

References

Academia Brasileira de Letras (2001): Vocabulário Ortográfico da Língua Portuguesa. 6th. edition. https://www.academia.org.br/nossa-lingua/busca-no-vocabulario (last access: 04-04-2022).

Alves, I. M. (2009): Néologie et société mondialisée: implications morphologiques. In: Actes des 8è Journées Scientifiques du Réseau de chercheurs Lexicologie, Terminologie, Traduction. Paris, p. 367–373.

Alves, I. M. (2007): The social aspects of Brazilian Portuguese neologisms. Os aspectos sociais dos neologismos do português brasileiro. Dictionaries (Terre Haute, Ind.) 28, pp. 149–156.

Alves, I. M. (2004): A unidade lexical neológica: do histórico-social ao morfológico. In: Aparecida Negri Isquerdo; Maria da Graça Krieger (Orgs.). As ciências do léxico: lexicologia, lexicografia, terminologia. Campo Grande: Editora UFMS, vol. 2, pp. 77–88.

Alves, I. M./Maroneze, B. O. (2021): The presence of Brazilian neologisms in dictionaries. International Journal of Lexicography 34, pp. 315–335.

Aulete Digital. Lexikon Editora Digital. https://aulete.com.br/ (last access: 04-04-2022)

Bauer, L. (2004): A glossary of morphology. Washington, DC.

Bizzocchi, A, (2021): Diário de um linguista. https://diariodeumlinguista.com/tag/fractocomposicao/ (last access: 04-04-2022).

Gonçalves, C.s A. (2016): Atuais tendências em formação de palavras. São Paulo.

Houaiss, A./Villar, M. de S.: Grande Dicionário Houaiss. https://houaiss.uol.com.br/ (last access: 04-04-2022).

Lino, M. T. R. da F. (1990): Observatório do português contemporâneo. In: Actas do Colóquio de Lexicologia e Lexicografia. Lisboa, pp. 28–33.

Michaelis: Dicionário Brasileiro da Língua Portuguesa. Ed. Melhoramentos. https://michaelis.uol.com.br/ (last access: 04-04-2022).

Sablayrolles, J.-F. (2017): Les néologismes. Créer des mots français aujourd´hui. Paris.

Takeda, T.: A preocupação com o meio ambiente nas últimas décadas. https://www.jurisway.org.br/v2/dhall.asp?id_dh=1762 (last access: 04-04-2022).

Tournier, J. (1985): Introduction descriptive à la lexicogénétique de l´anglais contemporain. Paris/Geneva.

Villela, P. (2015): As dimensões internacionais das políticas brasileiras de combate ao tráfico de drogas na década de 1990. Dissertação (Mestrado). PUCSP (Programa de Pós-Graduação San Tiago Dantas (UNESP, UNICAMP, PUCSP). https://www.funag.gov.br/ipri/btd/index.php/10-dissertacoes/1671-as-dimensoes-internacionais-das-politicas-brasileiras-de-combate-ao-trafico-de-drogas-na-decada-de-1990 (last access: 04-04-2022)

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Contact information

Ieda Maria Alves
Universidade de São Paulo
iemalves@usp.br

Bruno Maroneze
Universidade de São Paulo
maronezebruno@yahoo.com.br

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Jun Choi/Hae-Yun Jung

ON LOANS IN KOREAN NEW WORD FORMATION AND IN LEXICOGRAPHY

Abstract This study examines a list of 3,413 neologisms containing one or more borrowed item, which was compiled using the databases built by the Korean Neologism Investigation Project. Etymological aspects and morphological aspects are taken into consideration to show that, besides the overwhelming prevalence of English-based neologisms, particular loans from particular languages play a significant role in the prolific formation of Korean neologisms. Aspects of the lexicographic inclusion of loan-based neologisms demonstrate the need for Korean neologism and lexicography research to broaden its scopes in terms of methodology and attitudes, while also providing a glimpse of changes.

Keywords Neologisms; lexicography; loans; clipping; blending; word formation

1. Introduction

The objective of this paper is to understand what role loans play in Korean neologism formation and whether they weigh in the inclusion of loan-based neologisms in Korean dictionaries by analysing the neologisms that include at least one borrowed element. These neologisms were retrieved from the lists of all neologisms that were extracted from 2006 to 2019 within the framework of the Korean Neologism Investigation Project (a project funded and supervised by the National Institute of Korean Language) and in 2020 by the Centre for Korean Language Information Studies (Kyungpook National University). In section 2.1, an overview of the government-affiliated project is presented, along with an explanation of the methodology used to extract Korean neologisms.

By 'borrowed element' is meant any type of loans – whether a loanword, a loan-morpheme, or a clipped loanword to form blend neologisms. In this paper, loans are analysed according to three of their characteristics. First, they are examined from an etymological perspective (section 2.2). The language origins of borrowed items may range from European languages to Asian languages, but one language type that is not considered for loans is Traditional Chinese, or Hanja, as opposed to Simplified Chinese. Hanja-based words (also called Sino-Korean words) are regarded as fully Korean, even though they are distinguished from native Korean words (sometimes referred to as pure Korean words). Second, loan-based neologisms as well as Korean neologisms are analysed from a morphological perspective (section 3), that is, based on Korean word formation processes (3.1) and with regard to their potential productivity (3.2). Lastly, they are discussed from a lexicographic perspective (section 4), not only in terms of statistics but also in terms of the attitudes of the Korean academia towards loanwords and neologisms.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

2. Methodology

2.1 The Korean Neologism Investigation Project

Affiliated to the South Korean Ministry of Culture and Tourism since 1991, the National Institute of Korean Language (NIKL) undertook the task of collecting and analysing new words in Korean language from 1994 to 2019 under the project name 'Korean Neologism Investigation Project' (KNIP). KNIP was carried out on a yearly basis by the research team from a research centre or a university, who produced annual reports, for the most part available on the NIKL website[1]. The reports present the neologisms in order of high frequency as well as in alphabetical order, and provide crucial information for neologism and dictionary research. Indeed, they describe each neologism in lexicographic style (i.e., in terms of pronunciation, etymology, part-of-speech, domain when applicable, definition, and examples) and indicate not only the date of first occurrence but also frequencies of the neologisms

The project was temporarily discontinued in 2011 and resumed in 2012 with the Centre for Korean Language Information Studies at Kyungpook National University (KNU) carrying out the project under the supervision of NIKL until it came to an end in 2019. In 2020, the investigation into Korean neologisms was conducted independently by the abovementioned KNU Centre, using the same framework and criteria set by NIKL for the consistency of the data[2] (Nam et al. 2021).

In the early stages of KNIP, new words[3] were manually retrieved from printed newspaper articles and news broadcasting scripts. The development of the Internet and computing tools allowed a number of methodological improvements from the early 2000s onwards, including the distinction between neologisms proper and words that are simply not included in the *Standard Korean Language Dictionary* (SKLD), the expansion of source texts with ever-increasing online media[4] and, from 2005, the construction and use of a Web-based corpus to extract the neologisms automatically (in addition to the manual extraction). From 2012 onwards, neologism candidates have been retrieved using a Web-based neologism extractor based on whether or not a candidate word is represented in the online dictionary *Urimalsaem*[5,6]. The list of neologism candidates is then manually checked by researchers to narrow down the candidates to neologism headword candidates.

1 https://www.korean.go.kr/front/reportData/reportDataList.do?mn_id=207.

2 The report on the 2020 neologisms is not available in the NIKL website as the investigation was not carried out as part of the government-funded KNIP but has been published by Hankwukmwunhwasa.

3 We specifically use the term 'new word' and not 'neologism' here as the project aimed to retrieve any words that were not included in the *Standard Korean Language Dictionary* published by NIKL, whether neologisms or not.

4 The online news articles used for the project are those provided by the Naver News portal (https://news.naver.com/).

5 While the macrostructure of *Urimalsaem* is based on *SKLD* and administered by NIKL, it is nonetheless much bigger than *SKLD*. As an online dictionary, *Urimalsaem* has no printing limitations and also allows users to suggest headword candidates.

6 For more details on the methodological changes that were brought to the Korean Neologism Investigation Project, see Nam/Lee/Jung (2020, pp. 107–110); Choi (2020, p. 153).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

2.2 The neologisms under study

The present study focuses on the years 2006 to 2020[7] – that is, after the project started to use the corpus methodology and the retrieval of neologisms thus became more systematized. More specifically, the study targets the neologisms that are either full loanwords or partly borrowed. During this time frame, there were in total 6,554 neologisms collected, from which we extracted a list of 3,413 neologisms with at least one borrowed element by excluding all neologisms that are composed of solely native Korean and Hanja characters. Loan-based neologisms represent a little over 52% of the Korean neologisms collected in the past fifteen years or so. Table 1 shows the ratio of such neologisms for each year of the time frame and Table 2 the ratio of neologisms comprising at least one borrowed element from a given language, presented in order of high frequency.

Year	Number of neologisms	Number of loan-based neologisms	Percentage
2006	530	254	47.92
2007	702	369	52.66
2008	475	167	35.16
2009	588	219	37.24
2010	368	170	46.20
2012	511	298	58.32
2013	488	282	57.79
2014	339	212	62.54
2015	285	166	58.25
2016	649	349	53.78
2017	396	237	59.85
2018	460	291	63.26
2019	358	212	59.22
2020	405	187	46.17
Total	6,554	3413	52.07

Table 1: Ratio of loan-based neologisms per year

Borrowed language	Number of loan-based neologisms with at least 1 element from the borrowed language	Percentage
English (EN)	3,235	94.78
French (FR)	97	2.84
Japanese (JA)	70	2.05
Italian (IT)	52	1.52
Chinese (ZH)	22	0.64
German (DE)	22	0.64

[7] Except for 2011, where there is no data.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Borrowed language	Number of loan-based neologisms with at least 1 element from the borrowed language	Percentage
Latin (LA)	14	0.41
Greek (EL)	10	0.29
Spanish (ES)	5	0.14
Danish (DA)	4	0.11
Russian (RU)	3	0.08
Hindi (HI)	2	0.05
Pashto (PS)	2	0.05
Indonesian (ID)	1	0.02
Portuguese (PT)	1	0.02
Sanskrit (SA)	1	0.02
Turkish (TR)	1	0.02

Table 2: Ratio of loan-based neologisms per borrowed language

Although English is clearly and by far the predominant language as regards Korean neologisms, neologisms may borrow from a variety of languages. More importantly, they may combine elements from several languages. Thus, they can be divided into neologisms borrowing from a single language (1a), hybrid neologisms composed of borrowed elements from multiple foreign origins (i. e., other than native Korean and Traditional Chinese) (1b), and hybrid neologisms composed of a borrowed element and a native Korean and/or Hanja element (1c).

(1) a. *kheyleynsia* (ES *querencia*); *teykacicum* (FR *dégagisme*); *koltu misu* (EN *gold miss*)
b. *hwuykeylaiphu* (DA *hygge* + EN *life*); *lamulliey* (JA *râ*[*men*] + FR [*som*]*melier*); *weyting alpa* (EN *wedding* + DE *Arbe*[*it*])
c. *lattey appa* (IT *latte* + Korean (KO) 'daddy'), *takkwuin* (EN *di*[*ary*] + KO *kkwu*[*minun*] 'decorating' + Hanja (HA) 'person') [8]

Hybrid neologisms are the most common neologisms with borrowed elements, counting 2,225 items and constituting almost two thirds of all loan-based neologisms. Among those, only 106 neologisms fall into the (1b) type of hybrids. In other words, 95.2% of hybrid neologisms include at least one element in native Korean or Sino-Korean. Neologisms consisting of a single borrowed language are not necessarily loanwords. As seen in (1a), *koltu misu* (EN *gold* + EN *miss*), which is composed of English words, falls into the category of Konglish neologisms, whereby English morphemes are borrowed to form words in English which do not exist in the English language. Instead, Konglish neologisms follow the semantic and cognitive patterns of Korean. The following section explains the processes and trends in Korean word formation.

[8] Romanization of Korean follows the Yale romanization system and literal translation is provided in single quotation marks where needed. Round brackets show the original word in case of loans; square brackets show elements that have been dropped in the word creation process.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

3. Loans and word formation in Korean

3.1 Word formation processes in Korean neologisms

From a morphological perspective, Korean words can be divided into two main categories, that is, simplex words and complex words. The formation of simplex neologisms encompasses the rather rare generation of purely native Korean or Sino-Korean forms[9] as well as the borrowing of a wordform from a foreign language. On the other hand, complex words are formed by combining either a root and an affix (derivatives), or two or more word stems (compounds), or two or more clipped words (blends). In practice, the lines delimitating these categories are not clear-cut. For some native Korean/Sino-Korean neologisms for example, it is unclear whether they are simplex or complex words. This is the case of the 2018 neologism *pposilaeki* 'little one', used to designate young or small and cute people or animals. It could be argued that the form *ppo-* is short for *ppoccak*, which is a dialectal form for *paccak* 'close(ly)' and has been widely used by netizens with the meaning of 'cute' when describing animals or talking of Korean pop idols online. However, the word has been categorized as a simplex form, as a semantic neologism of *pposilaeki* 'crumb' in a southwestern dialect by analogy with the netizens' use of *ppoccak*.

Regarding loanwords, their categorization is not always obvious depending on whether they are considered from the point of view of word formation or from the perspective of the resulting word form. As ten Hacken/Panocová put it, even "if the borrowed word is the result of a word formation rule in the original language, the word formation origin is lost in the receiving language", because "[w]ord formation rules are not borrowed" (ibid. 2020, p. 4), only the final product is. The authors illustrated this with the English simplex word but originally German compound *kindergarten*. There are such examples in Korean neologisms, such as the 2018 loanword *khulipthocaykhing*, from the English blend 'cryptojacking' (*crypto*[*currency*] + [*hi*]*jacking*) which is perceived as a simplex word in the formation process of Korean neologisms. Even if there are about twice as many loanwords as neologisms generated from native Korean or Sino-Korean characters, they still only constitute about 6.7% of the total loan-based neologisms, most of which being blends or compounds.

Before getting into complex neologisms, a new (minor) process of word formation has developed following the ever-increasing user-generated content platforms, which can be somewhat considered as a morphological 'anomaly'. This process consists of forming neologisms by replacing Korean characters[10] with other characters of similar shape, regardless of their semantic or phonetic similarities, as the two 2018 neologisms illustrate in (2).

(2) 댕댕미 *tayngtayngmi* 'someone as cute as a puppy' for KR 멍멍이 *mengmengi* 'doggy'; 네주얼 *neycwuel* 'visual' instead of the English loanword 비주얼 *picwuel* 'visual'

These neologisms can be categorized as graphic neologisms or 'pictorial representations' (Kim 2016). Example (2) also shows that this graphic word formation process can be applied to both native Korean or Sino-Korean words and loanwords.

[9] Within the scope of our study, there are only 94 such neologisms out of the total 6,554 neologisms formed from 2006 to 2020.

[10] The Korean script can be classified as a syllabic alphabet in that it does not consist of ideograms as in Chinese, but of alphabetic letters that are combined in square clusters to form a character (i. e. a syllable).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

3.2 Compounding and blending in Korean neologisms: the case of French-based neologisms

As just mentioned above, most Korean neologisms, including loan-based neologisms, are formed through compounding and blending. A little more than a third of the neologisms under study are blends (1247 items) and nearly 45% of them are compounds (1526 items). When examined according to the language origin, we can observe clear patterns emerging between both categories. To illustrate these patterns, the following analysis focuses on French-based compound and blend neologisms.

(3) a. Compound neologisms: *hompakhangsucok* (EN *home* + FR *vacances* + HA 'tribe')
b. Blend neologisms: *nuckhangsucok* (KO 'late' + FR [*va*]*cances* + HA 'tribe'); *kolkhangsucok* (EN *golf* + FR [*va*]*cances* + HA 'tribe'); *molkhangsu* (EN *ma*[*ll*] + FR [*va*]*cances*); *holkhangsucok* (KO 'alone' + FR [*va*]*cances* + HA *cok* 'tribe'); *khakhangsu* (FR *ca*[*fé*] + FR [*va*]*cances*); *phwulkhangsu* (EN *pool* + FR [*va*]*cances*)

The most striking feature of the French-based neologisms is the salience of the loanword *pakhangsu* (*vacances*), especially used as the clipped loan *-khangsu* ([*va*]*cances*) to form blend neologisms. While (3.a) shows the only example of a compound formed based on the loanword, (3.b) features only a handful of blends formed with the clipped loan. Table 3 presents the number and percentage of such blends per year.

Year	Number of neologisms containing *khangsu* ([*va*]*cances*)	Percentage
2010	2	3.7
2011	0	0
2012	1	1.9
2013	4	7.4
2014	0	0
2015	1	1.9
2016	0	0
2017	1	1.9
2018	3	5.5
2019	14	25.9
2020	2	3.7
Total	28	51.9

Table 3: Ratio of blend neologisms containing the clipped French loan *-khangsu* ([*va*]*cances*) to French-based blend neologisms per year

28 out of the 54 French-based blend neologisms include the clipped loan *-khangsu*. While the *-khangsu* neologisms seemed to have been particularly trendy in 2019, distribution across the remaining years is rather balanced, which leads us to think that this particular loan will most likely continue to be used in the future. In fact, the loanword *pakhangsu* (*vacances*) and its clipped version have long been used to create Korean neologisms, many of which were included in *Urimalsaem*: twelve *pakhangsu* neologisms and fifteen *-khangsu* neologisms, including four from our list. The case of the French loan *-khangsu* can be ex-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

tended to a few other donning languages, especially where the number of neologisms is higher[11], as illustrated in Table 4.

Language	Number of loan-based neologisms	Most productive loan	Number of neologisms with most productive loan	Examples
Japanese	70	*otaku* 'geek'	2 compounds; 24 blends	*otekcil* (JA *otak*[*u*] + KO 'attitude'); *sengtek* (KO 'successful' + JA [*o*]*tak*[*u*])
Italian	52	paparazzi	3 compounds; 24 blends	*phaynphalachi* (EN *fan* + IT [*pa*]*parazzi*); *kyenphalachi* (HA 'muzzle' + IT [*pa*] *parazzi*)
Chinese	22	*mala* (spicy seasoning)	1 compound; 3 blends	*malamama* (ZH *mala* + ZH 'mummy')
German	22	Arbeit	1 compound; 7 blends	*alpaleylla* (DE *Arbe*[*it*] + EN [*Cinde*[*rella*]]
Latin	14	*homo* + attribute[12]	3 compounds; 5 blends	*homo cheyekhwusu* (LA *homo* + EN *chair* + LA –[Australopithe]cus)

Table 4: Most productive loans per language with higher number of neologisms containing the most productive loan

Table 4 shows that some loans yield many neologisms. It also confirms that their productivity is more prominent in blending when used in their clipped forms. This means that these particular loans are well established in the mental lexicon of Korean language speakers.

4. Lexicographic representation of and attitudes towards Korean loan-based neologisms

Just as *vacances*, such productive loanwords as *otaku*, *paparazzi*, *Arbeit*, and *homo*, as well as a number of the compound and blend neologisms they produced, have been included in *Urimalsaem*. However, they are not equally represented, as shown in Table 5.

[11] We leave alone the case of English, which is overwhelmingly higher and thereby would present many cases of productive (clipped) loanwords.

[12] That is, neologisms that imitate forms such as 'homo Australopithecus' or 'homo sapiens' to designate people living a certain lifestyle. For example, homo cheyekhwusu 'homo chaircus' refers to those who spend their day sitting, such as office workers or academics.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Loanword	Number of related neologisms	Number of related neologisms included in *Urimalsaem*	Number of other related neologisms included in *Urimalsaem* which are not from our list
pakhangsu (FR *vacances*)	28	4	23
othakhwu (JA *otaku* 'geek')	26	15	–
phaphalachi (IT *paparazzi*)	27	5	29
alupaithu, *alpa*[13] (DE *Arbeit*)	8	4	29
homo + attribute (LA *homo*)	8	7	18

Table 5: Representation of highly productive loans in *Urimalsaem*

While very few neologisms from our list made it to the dictionary, quite surprisingly, many other neologisms formed with the loanwords (compounded or blended) from Table 5 have been added to *Urimalsaem* but are not from the list of neologisms collected within the scheme of the KNIP. On the one hand, this highlights the limits of neologism extraction from the sole genre of 'news media'. On the other hand, it implies that these 'other' neologisms that eluded the project have been probably included following the dictionary users' suggestions. Indeed, *Urimalsaem* allows users to participate in the making of the dictionary and suggest new headwords with their definitions under a separate tab. Users' suggestions are then reviewed by language experts and potentially added if they have lexicographic value.

In addition to the inconsistent representation of these neological 'families', only FR *vacances* and DE *Arbeit* are represented in *SKLD*, moreover only in their full, unaltered forms. Although both dictionaries are managed by the NIKL and considered as language authorities, they also have different characteristics. Unlike *SKLD* that started off as a print dictionary that has been digitalized, *Urimalsaem* is solely an online dictionary which has some characteristic features of online content. For example, as mentioned earlier, it allows user-generated content, although still supervised by experts. Thus, *Urimalsaem* is more inclusive – at least on paper – and its macrostructure, albeit based on *SKLD*, grows at a faster rate than *SKLD* which is more of a traditional dictionary, that is, prescriptive and more conservative towards neologisms, and particularly loan-based neologisms.

In the mid-1970s, the government undertook the task to 'purify' the Korean language. The main objective of the task was to refine Korean by replacing improper or dialectal words by correct, standard words, complicated words or expressions by simpler ones, and words of foreign origins by native Korean words (Kim 2019; Seo 2019). *SKLD* and *Urimalsaem* are government-affiliated dictionaries and may reflect some aspects of the language purification policy. In the case of *SKLD*, it is shown from its macrostructure itself. In the latest statistical report on the dictionary content, which is accessible on the *SKLD* website[14], it ap-

[13] *Alpa* is the shortened form of *alupaithu*.

[14] https://stdict.korean.go.kr/statistic/dicStat.do#static_menu3_3.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

pears that words of foreign origin constitute a mere 5.6% of the headwords and hybrids account for 20.5%; Korean words, including native Korean (20.9%) and Sino-Korean (53%), make up the vast majority of the macrostructure. The low ratio of foreign words together with hybrids (26.1%), may be explained by the dictionary's passivity before neologisms in general, and loan-based neologisms in particular. As a matter of fact, one of the rare neologisms that were included in *SKLD* in the 2000s is *colipep* 'recipe', which was actually presented as a Korean replacement to the English loanword *leysiphi* 'recipe'.

As for *Urimalsaem*, it does accept more neologisms and loans; however, it has not neglected its role of language prescription. Example 4 shows a couple of cases of 'normative information' in loanword entries.

(4) a. *pakhangsu* (FR *vacances*):
Purification (notice for the correction of daily life terms (Ministry of Culture and Sports Notice No. 1996-13, March 23, 1996))
Instead of '*vacances*', use the refined terms of *yelum hyuka* (KO 'summer' + HA 'vacation') or *hyuka* (HA 'vacation') if possible.

b. *alupaithu* (DE Arbeit):
Purification (notice for the correction of daily life terms (Ministry of Culture and Sports Notice No. 1996-13, March 23, 1996))
'*pwuep*' (HA 'part-time job) can be used along '*Arbeit*'

It is nonetheless safe to say that many normative forms of the Ministry of Culture and Sports fall into oblivion.

5. Conclusive remark

As for a conclusion, we propose to have a last look into our list of loan-based neologisms and check how, overall, they are represented in *Urimalsaem* and how they compare with native Korean and Sino-Korean neologisms.

Year	Total neologisms	Total neologisms included in *Urimalsaem*	Number of native Korean and Hanja neologisms represented	%	Number of loan-based neologisms represented	%
2006	530	91	58	63.7	33	36.3
2007	702	148	70	47.3	78	52.7
2008	475	85	56	65.9	29	34.1
2009	588	97	70	72.2	27	27.8
2010	368	17	8	47.1	9	52.9
2012	511	178	73	41.0	105	59.0
2013	488	152	56	36.8	96	63.2
2014	339	196	85	43.4	111	56.6
2015	285	206	85	41.3	121	58.7
2016	649	302	143	47.4	159	52.6
2017	396	211	76	36.0	135	64.0

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Year	Total neologisms	Total neologisms included in *Urimalsaem*	Number of native Korean and Hanja neologisms represented	%	Number of loan-based neologisms represented	%
2018	460	201	72	35.8	129	64.2
2019	358	20	8	40.0	12	60.0
2020	405	8	3	37.5	5	62.5
Total	6,554	1,912	863	45.1	1,049	54.9

Table 6: Lexicographic representation of native Korean and Hanja neologisms against loan-based neologisms per year

Table 6 allows us to take a look at the big picture. The inclusion of loan-based neologisms can be divided into three main stages. Until 2009, there were generally fewer loan-based neologisms included in the dictionary than native Korean and Hanja neologisms, regardless of whether their ratio was higher. Then, in the first half of the 2010s, they seemed to gain ground in the lexicographic race. Finally, since 2017, even though fewer neologisms have been added to the dictionary, there tend to be twice as much inclusion of loan-based neologisms as native Korean and Hanja neologisms. Despite the normative attitude of Korean dictionaries and the efforts of language policies to minimize the impact of loanwords on the Korean language, it is the language speaker who ultimately shapes the language by creating new words and choosing the words to use and giving momentum to loan-based neologisms.

References

Choi, J. (2020): On the methodology for the detection of Korean neologisms. In: Language Facts and Perspectives 51, pp. 151–172.

ten Hacken, P./Koliopoulou, M. (2020): Dictionaries, neologisms, and linguistic purism. In: International Journal of Lexicography 33 (2), pp. 127–134.

ten Hacken, P./Panacová, R. (eds.) (2020): The interaction of borrowing and word formation. Edinburgh.

Kim, E. J. (2016): Anglicized Korean neologisms of the new millennium: an overview. In: English Today 32 (3), pp. 52–60.

Kim, J. (2019): The flows and changes of Korean language purification policy. In: Culture and Convergence 41 (4), pp. 1279–1303.

Nam, K./Lee, S./Jung, H. (2020): The Korean Neologism Investigation Project: current status and key issues. In: Dictionaries: Journal of the Dictionary Society of North America 41 (1), pp. 105–129.

Nam, K./Lee, S./Choi, J./Seo, E./Kang, H./Baek, M./Jeong, H./Kim, H./An, J. (2021): Neologisms of 2020. A new language of the COVID-19 pandemic. Seoul.

Seo, H. (2019): On the meaning correspondence of refining words. In: Korean Semantics 64, pp. 131–153.

Standard Korean Language Dictionary (2002). https://stdict.korean.go.kr/main/main.do#main_logo_id (last access: 27-03-2022).

Urimalsaem (2016). https://opendict.korean.go.kr/main (last access: 27-03-2022).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Contact information

Jun Choi
Kyungpook National University
c-juni@hanmail.net

Hae-Yun Jung
Kyungpook National University
haeyun.jung.22@gmail.com

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Lars Trap-Jensen/Henrik Lorentzen

RECENT NEOLOGISMS PROVOKED BY COVID-19 IN THE DANISH LANGUAGE AND IN THE DANISH DICTIONARY

Abstract Inspired by GWLN 3, we take a look at the new words, meanings, and expressions that have been created during or promoted by the COVID-19 pandemic. The pandemic provides a rare opportunity to follow the rise, spread, and integration of words and expressions in a language that may serve as an illustration of how linguistic innovation in general works. Relevant words were selected from various lists, notably monthly and annual lists of prominent words attested in the corpus of The Danish Dictionary. Analysis of these lists gives an insight into the number of words that stand out month by month and what kinds of words are involved, both in terms of morphological type and of semantic category, with special attention given to neologisms. Finally, we discuss the criteria for selecting which words to include in the dictionary. With this study, Danish is added to the list of languages covered in the GWLN series on COVID-19 neologisms.

Keywords COVID-19; detecting neologisms; corpus-based; temporal dimension; The Danish Dictionary

1. Background: dictionary and corpus

The Danish Dictionary is a descriptive dictionary that uses its own corpus as the primary source of new lemmas. The dictionary was first published in six printed volumes 2003-2005 and later converted into an internet dictionary. Since 2009, it has been available at the site https://ordnet.dk, with all subsequent additions and revisions being published online only. No new printed edition is foreseen. The publisher is the Society for Danish Language and Literature, an independent research and editing institution sponsored in part by the Danish Ministry of Culture, in part by private foundations.

As a descriptive dictionary, The Danish Dictionary developed its own corpus resources in the 1990s and has continued to collect text material since then. As of 2022, this means that we have a monitor corpus with texts from c. 1983 onwards, containing 1.1 billion words and growing every month so that the editors can monitor linguistic development over time (https://korpus.dsl.dk/documentation). One way of doing this is looking at lists of frequent and prominent words in a particular month or a particular year, a feature that was implemented into the corpus querying tool (CoREST, developed in-house) and has been available to the editors from 2016.[1]

2. Methodology: detecting COVID-19 neologisms

To find COVID-19 relevant words, we browsed through the monthly and annual lists of prominent and frequent words from March 2020 until January 2022 and extracted potential

[1] Frequent and prominent words are extracted by comparing texts from a particular period (a month or a year) with a reference corpus (all texts), using the measure Mutual Information. Subsequently, different filters are applied to the annual lists to yield widely used (high-frequency range) and prominent (lower frequency range) words respectively. Both the annual and monthly lists use a lower cut-off point to eliminate nonce occurrences.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

neologisms and other COVID-19 related words. The resulting list of candidates then served as input to further (manual) queries in the corpus – obvious examples being compounds and derivations with *COVID* and *corona*. In addition to this, we used various other word lists available to the editors via the corpus query tool: suggestions from the dictionary's users (a database of more than 30,000 posts), the editors' observations, word lists generated by individuals or public institutions (notably the Danish Language Council), etc.

All in all, the corpus material and the word lists collected in this way include an inventory of approx. 900 COVID-19 related words.

It must be borne in mind, among other things, that not all words related to COVID-19 are neologisms and, conversely, not all neologisms are about COVID-19. The two sets partially overlap, and the inventory of the intersection set is not easy (or even possible) to delimit with precision. First, the inventory is probably too short as we have undoubtedly overlooked relevant words in the process, not least because the low-frequency words below a certain threshold were excluded from the monthly list (useful for editors in search for good lemma candidates, less fortunate for this particular task). Second, the inventory is probably too long because there is no objective way of telling what makes a word COVID-19 related. Is an expression like *video meeting* COVID-19 related? Probably yes, as the pandemic has boosted the use of both the concept and the expression enormously. On the other hand, no it isn't, as both existed before the lockdown of March and April 2020 and could, in principle, end up on the list for different reasons. One should also be careful about what is meant by the term *neologism*. In this context, we use it to include words and expressions that at a certain point in time could not be substantiated earlier, as well as pre-existing words and expressions that have gained new meaning or new or increased use (following Agazzi 2015, p. 7).[2] For practical reasons, a cut-off point of 30 years is used in The Danish Dictionary (somewhat arbitrarily, cf. Trap-Jensen 2020), and words younger than this are regarded as neologisms. Even this definition is not entirely without its problems. For a word to be counted as a neologism, it must be part of the language, and that in turn requires some degree of integration into the language that separates it from nonce occurrences.

However, we still think it is worthwhile using and analysing the list of candidates – with the reservation that the list is neither authoritative nor exhaustive.

3. COVID-19 relatedness

To decide if a word is related to COVID-19 is no easy task – for obvious reasons: the pandemic caused an all-embracing crisis which had at its centre the coronavirus and the COVID-19 disease itself but with far-reaching implications for almost all aspects of society: legislation, economy, politics, business, leisure, education, etc. Of course, this is reflected in the words we use to talk about these topics. The resulting lockdown sent people home to an entirely new virtual experience with Zoom meetings, online classrooms, and regular televised press conferences. New legislation was introduced to regulate social life and compensate shops, bars, and other businesses that suffered from the lockdown.

We mention this simply to stress how complicated it is to judge whether a word is COVID-19 relevant by just looking at it in isolation. It requires world knowledge to tell that words like

[2] Our English translation of the Swedish original: "dels ord och uttryck som vid en viss tidpunkt inte kunnat beläggas tidigare, dels redan existerande 'äldre' ord och uttryck som fått ny användning eller ny betydelse eller ökad användning."

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

afspritning ('cleaning with alcohol gel'), *omsætningsfald* ('turnover decline'), and *forsamlingsforbud* ('ban on public gatherings') are COVID-19 related through the chain of measures that were taken following the health recommendations and political actions. It also implies that COVID-19 related words cannot be extracted automatically by a computer – even two humans may not come up with the same result if asked to select relevant words from a list.

4. The temporal dimension: monthly lists

In themselves, the lists of "words of the month" are interesting to study (cf. Salazar/Wild 2021). The sheer number of words that appears on a monthly list tells a story of its own. Each word is there because it is either (a) a new word of a certain frequency, or (b) an established word that is used with above-normal frequency for that month. In other words, an uneventful month will result in a list with fewer words than a more eventful one, and vice versa.[3] With that in mind, it is significant that the lists for the first two months of the pandemic in Denmark, March and April 2020, stand out as displaying by far the highest number of words: where the average number of words for a month during the period considered here (March 2020 until January 2022) is 51, March shows 185 and April 174.

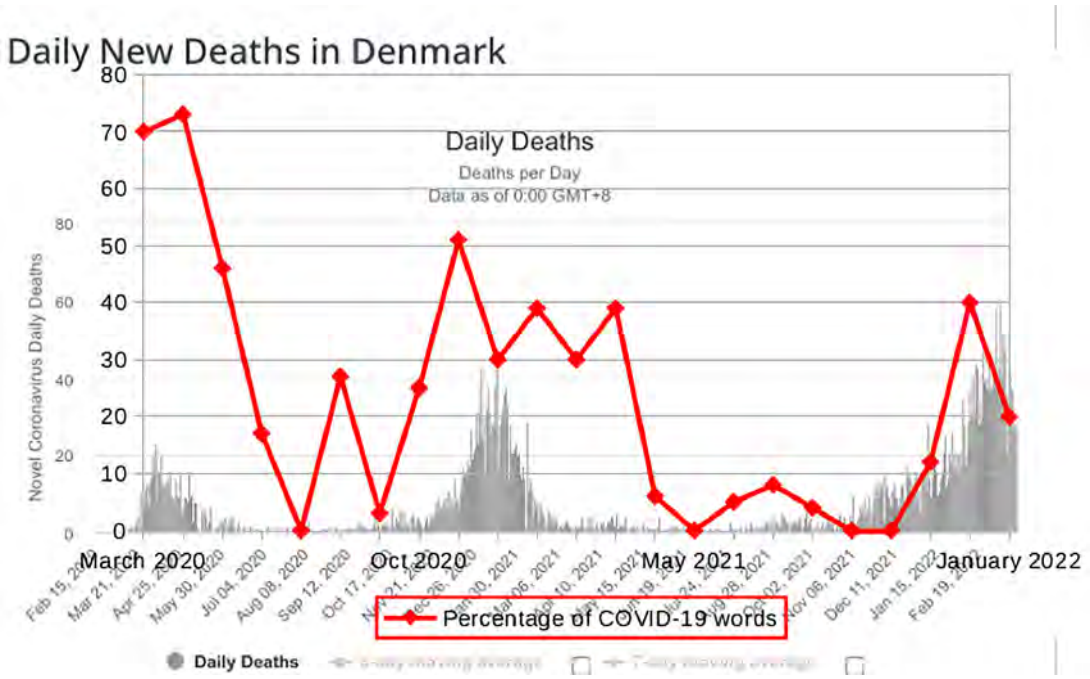


Fig. 1: The percentage of COVID-19 relevant words on monthly lists compared with COVID-19 deaths as provided by Johns Hopkins (https://www.worldometers.info/coronavirus/country/denmark/)

Moreover, the proportion of words that are COVID-19 relevant is also the highest in the period: 70 and 73 per cent respectively (bearing in mind the somewhat subjective nature of judging relevance).

[3] Apart from the COVID-19 words, other significant events can also be traced: the Tokyo Olympics in July and August 2021 with words like *medaljehåb* ('medal hope'), *holdforfølgelsesløb* ('team pursuit') and *skeetskytte* ('skeet shooter'), and the local elections in Denmark November 2021 promoting *borgmesterpost* ('mayor's office'), *spidskandidat* ('front runner') and *vælgerlussing* ('electoral defeat').

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

For what it is worth, the quantitative peaks of the monthly lists coincide neatly with the pattern of the pandemic itself as demonstrated in Figure 1. Here we have used the proportion of COVID-19 relevant words but the picture remains unchanged even if absolute numbers are used.

If we zoom in to take a closer look at the contents of the lists, we can follow the development of the pandemic. In March and April 2020 the top of the lists reflects the societal crisis with lockdown, pressure on hospitals etc. through words like: *respirator* ('ventilator'), *værnemiddel* ('personal protective equipment, PPE'), *smittekæde* ('infection chain'); *hjemmekarantæne* ('home quarantine'), *fjernundervisning* ('distance teaching'); *hamstre* ('stockpile'), *toiletpapir* ('toilet paper'), *hjælpepakke* ('aid package'), *likviditet* ('liquidity, cashflow'). The highest scoring of the 185 words on the list for March 2020 are shown in Figure 2.

Månedens ord | 2020-03

Query	Score
corona	120089
coronavirus	77654
coronasmitte	25868
covid	24483
pandemi	22856
hjælpepakke	21147
virus	19187
smittespredning	15638
håndsprit	12624
coronaepidemi	10996
smitte	9695
nedlukning	9254
epidemi	7945
coronasituation	5891
respirator	5110
hamstre	5057
værnemiddel	5032
spritte	5017
smittefare	5010
smittekæde	4955
hjemmekarantæne	4883
hamstring	4735
grænselukning	4646
smittet	4434
smitterisiko	4281
coronaudbrud	4075

corona
corona virus
corona infection
COVID
pandemic
aid package
virus
spread of infection
hand sanitizer
corona epidemic
infection
lockdown
epidemic
corona situation
ventilator
stockpile
personal protective equipment, PPE
wash with alcohol gel
infection danger
infection chain
home quarantine
stockpiling
border closure
infected
infection risk
corona outbreak

Fig. 2: Words of the month: most frequent from March 2020 (English translations right)

The picture changes gradually in May 2020 when the most prominent word was *genåbning* ('reopening'). Corona words are still frequent, but now also in compounds like *coronapause*, reminding us that the lockdown is temporary and that a slow return to more normal conditions is beginning. Throughout June and July 2020, COVID-19 related words are not particularly salient in the lists, only to return to the news media again in October and November when a virus variant was detected first in mink and shortly after in humans. Fearing a resurgence of the pandemic the Danish government decided to order all the country's mink culled and buried, a decision that caused a lot of public debate as the legal basis of the decision was not in place. As a consequence, words related to the mink industry dominate the monthly lists, and words more narrowly connected with COVID-19 return (e. g. *mutere* 'mutate' and *mutation*) as well as words related to the development of vaccines, e. g. *coronavac-*

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

cine, vaccinedosis, vaccinationsplan, vaccinationscenter. During this second wave of the epidemic in Denmark, new words were introduced: *lyntest* ('immediate test'), *hurtigtest* ('speedy test'), *kviktest* ('quick test'), *coronapas* ('corona pass'), and the general feeling of 2020 as a special year marked by COVID-19 is expressed by words like *coronaår* ('corona year'), *annus horribilis* and *coronatræthed* ('corona fatigue'). After a long and relatively strict lockdown in the first months of 2021, the virus returned in December and could be observed in the monthly lists through words such as *omikron*, *omikronvariant*, *boosterstik* ('booster shot'), *revaccination*, and a change in official testing policy towards fewer PCR tests can be observed in the occurrence of the words and concept of *hjemmetest* ('home test') and *selvtest* ('self-test').

5. Lexical types of COVID-19 words and neologisms

If we look at the total amount of words from all the lists available, they may be divided into different general groups based on the linguistic type or their status in relation to the pandemic.

5.1 New coinages

First, we have genuinely new coinages that have never been used in the language before; most obvious is the term *COVID-19* itself (named by WHO in February 2020), but also *coronatest, coronaprøve* ('corona test'). The terms *corona* and *coronavirus* had existed in specialised language before 2020, but entered and dominated the general language, and they were included alongside *COVID-19* in the dictionary updates of June and November 2020. Another example of this came with the vaccines, i. e. the word *coronapas* ('corona pass'), which did not exist until 2021, and it will be included in the dictionary update of June this year together with *vaccinepas* ('vaccine pass'), a word that was attested first in 2014 but was boosted immensely in the recent two years.

5.2 Pre-existing 'pandemic' words

Second, we have seen pre-existing words revive and spread from a dormant state, e. g. words that were used in connection with previous pandemics: *selvisolering* ('self-isolation'), *hjemmekarantæne* ('home quarantine'), and *pandemisk* ('pandemic', adjective). Several words of this type were included in the updates of 2020.

5.3 Specialised words

Third, a large number of words that otherwise belong to specialist domains like virology and medicine, have extended their use into the common language as these have become the subject of our news feeds and everyday discussions, e. g. *asymptomatisk* ('asymptomatic'), *samfundssmitte* ('community spread'), *flokimmunitet* ('herd immunity'), *PCR-test*, and *superspreder* ('super spreader'). We have added around 30 words of this type, including also terms that are slightly less connected to specialised language but have nevertheless become common in everyday language during the different phases of the pandemic. Examples of this are *podepind* ('swab'), *boostervaccine*, *smittetal* ('infection rate'), *testkit*, *udrulning* ('roll-out' of vaccines). It can be argued that a general-language dictionary is not obliged to cover spe-

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

cialist language, but the boundaries between those domains are not fixed, and the corona crisis is an excellent example of a situation where the general public all of a sudden need information about specialist subjects.

5.4 Words related to the crisis in general

A fourth group consists of words related to the corona crisis in a broader sense. Some of them pertain to the society being locked down, e.g. *videomøde* ('video meeting'), *forsigtighedsprincip* ('precautionary principle'), and *rejserestriktion* ('travel restriction'). Another group of new words is promoted by the political life and all the political measures and appearances we have witnessed. Prominent new words that found their way into the dictionary from this group include *doorstep* ('short press conference') and *hastelov* ('law passed in a hurry'). The word *landsmoder* ('mother of the country') is quite notable in a Danish context. It can be attested sporadically over the last decades, but it goes sky-high in 2020 and later, the obvious reason being the dominant role and position of Danish Prime Minister Mette Frederiksen, who has become very popular – and also widely criticised during all of the corona crisis. A third group of general crisis words that have been added to the dictionary shows that the crisis is also of an economic nature: *helikopterpenge* ('helicopter money', distributed to boost consumer spending), *forsyningsproblem* ('supply problem').

6. COVID-19 words in the dictionary

The pandemic has generated far more neologisms than are reflected in the dictionary. Most likely, many of the new words will not stay in the language after the crisis. Highly professional words will again be confined to the professional domains and return to their dormant state in the general language. Nevertheless, it is interesting to watch how easily these words spread into everyday language when the need arises. Likewise, the wealth of nonce and slang words on the lists pay testimony to the way we cope with the COVID-19 crisis: *krammerat*, a blend of *kramme* 'hug' and *kammerat* 'friend, buddy', *maskne*, a blend of *maske* '(face) mask' and *akne* 'acne', and *corontæne*, a blend of *corona* and *karantæne* 'quarantine', are in all likelihood not going to stay very long in the language, and the same is true of the large number of *corona* compounds created over the last two years: *coronahår* ('corona hair' that you develop because hairdressers are closed), *coronakilo* ('corona kilo', referring to weight gain due to a lack of exercise when working from home) and *coronaturist* ('corona tourist' referring to city dwellers that flee to the countryside in an attempt to avoid the virus). They are probably just momentary occasionalisms that will be forgotten again after the crisis but they are important to note as evidence of how humour can be used as a strategy to deal with the crisis. But it does not mean that they should be included in a dictionary of the common language.

Until now, we have added 81 COVID-19 related entries to The Danish Dictionary. The editors try to assess the long-term durability of candidate words, a task that is not easy when you witness history in the making. As swift decisions must be made, it follows that some of the entries are exempted from the period of three years that under normal circumstances is used as a qualifying period before a neologism is included.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

7. Discussion and conclusions

There are several lessons to learn from the account of COVID-19 related words: We saw that the tumultuous events of March and April 2020 produced an unusually high number of words, new words, or words used with above-normal frequency. We may interpret this as a sign of language's ability to adapt quickly to new circumstances: we use or create the vocabulary that is necessary to cope with and communicate about the new situation in which we suddenly find ourselves. It is also reasonable to see the high percentage of COVID-19 related words as an expression of the all-embracing nature of the pandemic as experienced in the period.

We have not checked each word on the inventory (mentioned in section 2) separately to find the first recorded instance, but a qualified guess is that at most 10 per cent of the words are neologisms in the strict sense,[4] and even among these, the first recorded instance is likely to refer to events before the pandemic. But what is more striking, is how fast we adapt and become familiar with the technical vocabulary of immunology, social medicine, and neighbouring disciplines. One of the true neologisms on the list, *mandagsvirolog* ('Monday virologist'), is itself a nice – and witty – example of this inclination. It was created in analogy with the existing concept of *mandagstræner* ('Monday coach'), referring to football supporters who know everything their favourite team should have done – the day after they played. During the pandemic, we have all become experts in medicine and confidently drop words like *antigentest*, *incidensrate* ('incidence rate'), *komorbiditet* ('comorbidity'), and *kontakttal* ('reproduction rate, R rate'), words that most of us probably never had heard before the pandemic.

That is a reminder that to the individual it does not matter if a word has been used in a special field before. If he or she hears the word for the first time, it is as much a neologism to them as the true neologisms that nobody has heard before.

References

Agazzi, B. (2015): Neologisms in Swedish: blogg, fulbryt, pudla, rondellhund and other new editions from A to Ö [Nyord i svenskan: blogg, fulbryt, pudla, rondellhund och andra nytilskott från A till Ö]. Stockholm.

CoREST = Corpus Retrieval System & Tools (developed by Jørg Asmussen). https://korpus.dsl.dk/corest (last access: 20-03-2022).

Danish Language Council: List of corona words added to New Words in Danish. https://dsn.dk/ordboeger/nye-ord-i-dansk/et-coronaramt-ordforraad-nye-tilfoejelser-til-nye-ord-i-dansk/ (last access: 20-03-2022).

New words in Danish 1955 until today [Nye ord i dansk 1955 til i dag] (n. d.). The Danish Language Council. https://dsn.dk/ordboeger/nye-ord-i-dansk/om-nye-ord-i-dansk (last access: 20-03-2022).

Salazar, D./Wild, K. (2021): The *Oxford English Dictionary* and the language of Covid-19. 3rd Globalex Workshop on Lexicography and Neology, 2021.

[4] By this definition and with the cut-off point of 30 years mentioned earlier, only 14 of the COVID-19 related words included in the dictionary are true neologisms: *boostervaccine, coronapas*, *coronaprøve* ('corona test'), *covid-19*, *flokimmunitet* ('herd immunity'), *helikopterpenge* ('helicopter money'), *onlineundervisning* ('online teaching'), *PCR-test*, *solnedgangsklausul* ('sunset clause'), *stofmundbind* ('cloth facemask'), *superspreder* ('super spreader'), *testkit*, *vaccinationspas*, *vaccinepas*.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Swedish Language Council: lists of new words.
https://www.isof.se/lar-dig-mer/kunskapsbanker/lar-dig-mer-om-nyord/nyordslistan-2021 (last access: 20-03-2022), https://www.isof.se/lar-dig-mer/kunskapsbanker/lar-dig-mer-om-nyord/nyordslistor/nyordskronikor/2020-spraket-har-ocksa-fatt-corona (last access: 20-03-2022).

The Danish Dictionary [Den Danske Ordbog]. Society for Danish Language and Literature. https://ordnet.dk/ddo (last access: 20-03-2022).

Trap-Jensen, L. (2020): Language-internal neologisms and anglicisms: dealing with new words and expressions in The Danish Dictionary. In: Klosa-Kückelhaus, A./Kernerman, I. (eds.): Dictionaries. Journal of the Dictionary Society of North America. Volume 41, Issue 1, Special Issue: Global Viewpoints on Lexicography and Neologisms. Dictionary Society of North America 2020, pp. 11–25.

Contact information

Lars Trap-Jensen
Society for Danish Language and Literature
ltj@dsl.dk

Henrik Lorentzen
Society for Danish Language and Literature
hl@dsl.dk

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Gilles-Maurice de Schryver/Minah Nabirye

TOWARDS A MONITOR CORPUS FOR A BANTU LANGUAGE

A case study of neology detection in Lusoga

Abstract This paper looks at whether, after two decades of corpus building for the Bantu languages, the time is ripe to begin using monitor corpora. As a proof-of-concept, the usefulness of a Lusoga monitor corpus for lexicographic purposes, *in casu* for the detection of neologisms, both in terms of new words and new meanings, is investigated and found useful.

Keywords Monitor corpus; neology detection; new words; new meanings; Bantu; Lusoga

1. Corpus building for the Bantu languages

Corpus building efforts for the Bantu languages remain in their infancy, and not much has changed since the overview published in de Schryver/Prinsloo (2000) – with current corpus sizes typically anywhere between a million and five million tokens. These corpora have mainly been used for dictionary compilation, corpus linguistics, and NLP applications. The last collection of studies in the field of Bantu NLP is already a decade old (De Pauw et al. 2011), and includes studies for all languages from South Africa, as well as Swahili. Recent studies in Bantu corpus linguistics include Dom/de Schryver/Bostoen (2020) for Kisikongo, Kawalya/Bostoen/de Schryver (2021) for Luganda, and Misago/Nshimirimana/Tuyubahe (2021) for Kirundi. Examples of corpus-driven dictionaries compiled for Bantu languages over the past 15 years include de Schryver (2007) for Northern Sotho, de Schryver (2010) for Zulu, and de Schryver/Reynolds (2014) for Xhosa. In all these cases, one or more corpora were built, typically subdivided into a number of sub-corpora reflecting different time periods, genres, and/or topics. The majority of Bantu corpora to date are also 'raw', in that they have not been marked for parts of speech, nor been lemmatised. Also, no project so far has tried to build a 'monitor corpus' for a Bantu language, with which the changing language may be (semi-)automatically tracked (see e.g. Kosem et al. 2021; Kosem 2022). In the current study we attempt exactly that, and apply it to the detection of neologisms in Lusoga, with the aim of improving existing dictionaries for this language.

2. Corpus building for Lusoga

Lusoga is a Great Lakes Bantu language spoken in the Busoga Kingdom, in Eastern Uganda, by about three million people (UBOS 2016, p. 71). Despite a flurry of activity over the past two decades, it may still be classified as a predominantly oral language. During this period, the corpus building effort has been heroically carried forward by a single person (the second author of the present paper), as described in de Schryver/Nabirye (2018). Half a decade ago, the Lusoga corpus stood at a respectable 1.7m tokens (with an oral part of over half a million tokens, 541k more precisely), a corpus mainly used as 'the body of evidence' in writing the first corpus-based grammar of the language (Nabirye 2016). Corpus building continued unabated, and included a special focus on transcriptions of diverse oral data, to reach 3.0m

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

tokens in September 2019 (oral part: 786k; a selection and analysis of which was published in book form: Nabirye 2019).

Within the field of corpus building for the Bantu languages, the Lusoga corpus of 3.0m tokens was considered 'large enough', for it to be able to serve as a base for all future Lusoga studies.[1] Among the tests performed to judge whether or not the Lusoga corpus of 3.0m was also 'stable enough' to act as a reference corpus, stability tests similar to those described by Prinsloo/de Schryver (2001) for the Bantu languages Northern Sotho and Xitsonga were conducted.

Over the past two years, another half a million tokens were collected in addition, bringing the total size of the Lusoga corpus up to 3.5m tokens, nearly a million of them (910k) transcribed material. While it is still a raw corpus, the oral component corresponds to a massive 152 hours of audio recordings; the written component to about 16,000 pages of running text.

3. Towards a monitor corpus for Lusoga

The proof of a pudding is in the eating. It is one thing to judge that a corpus is large and stable enough to be used as a base and reference corpus; it is another entirely to also actually *use* it as such. One valuable use of such a corpus, if it does what it is supposed to do, is to act as a monitor corpus. In their standard textbook, McEnery/Hardie (2012, p. 246) define a 'monitor corpus' as: "A corpus that grows continually, with new texts being added over time so that the dataset continues to represent the most recent state of the language as well as earlier periods." Hanks (2003, p. 53) literally defines a 'dynamic' or so-called 'monitor corpus' in two words: "constantly growing". This may all be good and well and perhaps even feasible for big languages such as English, but for Bantu corpora with their typical modest sizes one can surely not simply keep adding material opportunistically, as one's corpus would lose all its balance and representativeness. As such, given that extreme care is taken to continuously balance out the genres and topics that are being added to the Lusoga corpus, so that it remains representative of both spoken and written Lusoga at all times, some earlier data are sometimes even removed before new material is added (see e.g. de Schryver/Nabirye 2018, § 3.3 vs. § 3.4). In a similar vein, and thus also for reasons of balance and representativeness, 1.6m tokens (of judicial material) have for instance always been kept separate from the main Kirundi corpus (Misago 2018, p. 38), or 2.0m tokens (of religious material) have always been kept separate from the main Luganda corpus (Kawalya 2017, § 5.2 vs. § 5.3 in Chapter 1). In this regard, Kilgarriff's characterisation of how to use monitor corpora for lexicographic purposes is probably more to the point:

> a long-standing vision is the 'monitor corpus', the moving corpus that lets the researcher explore language change objectively (Clear 1988, Janicivic and Walker 1997). The core method is to compare an older 'reference' corpus with an up-to-

[1] In order to put corpus sizes for Great Lakes Bantu languages in context, for the much larger language Kirundi (the national and official language of Burundi, spoken by 8 million people), three scholars contributed to the building of a Kirundi corpus to inform their respective PhDs: Mberamihigo (2014) built and used a Kirundi corpus of 1.9m tokens (oral part: 51k), Nshemezimana (2016) enlarged that to 2.2m tokens (oral part: 196k), and Misago (2018) reached 2.8m tokens (oral part: 418k). For Lusoga's bigger neighbour, Luganda (one of the national languages of Uganda, spoken by 6 million people), Kawalya (2017) built and used a corpus of 4 million tokens for his PhD. Both Kirundi and Luganda have a rich written tradition.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

> the-minute one to find words which are not already in the dictionary, and which are in the recent corpus but not in the older one. (Kilgarriff 2013, p. 81)[2]

The detection of 'new words' is not the only goal though, as dictionary compilers are also, and sometimes even more so, interested in the detection of new usages, and thus 'new meanings' (cf. Hanks 2002), of existing words:

> Monitor corpora are primarily of importance in lexicographic work [...] They enable lexicographers to trawl a stream of new texts looking for the occurrence of new words or for changing meanings of old words. (McEnery/Wilson 2001, p. 30)

Therefore, and in terms of methodology, we will now compare the additional 0.5m Lusoga material to the earlier 3.0m reference corpus. To do so, we make use of the KeyWords tool from WST (Scott 2019), which calculates the 'outstandingness' of each corpus type. The assumption is that we will be able to detect *new words* which entered the language, as well as *new meanings* for existing words. For the first we assume that we can obtain a limited list of new types in the additional 0.5m that were absent from the 3.0m. For the second we assume that a limited list of 'outstanding' types (specifically types used relatively more frequently over the past two years), will hint at extra usages and thus new meanings. While this exercise may seem trivial, it is not, as what one does not want is long lists of so-called 'new words' that are not new at all but were all simply missing from the 3.0m corpus, and/or 'new meanings' that are not new at all but were all simply not used in the 3.0m corpus. That a certain percentage was truly missing or not used is acceptable (no corpus, no matter its size, contains all the words in all its uses for any given language), but to usefully act as a monitor corpus, a useful percentage must be 'new' words and usages, or thus 'neologisms'. If neologisms are indeed detected in this way, lexicographers may also act upon those. That said, if this exercise is successful – in the sense that it results in meaningful data that can be acted upon by dictionary compilers – we can then consider the 3.5m corpus as the new reference and thus new monitor corpus.

4. The semi-automatic identification of neologisms in Lusoga

4.1 New words

The default settings of WST's KeyWords were used and, fair enough, a limited number of 55 keywords occurring in at least two of the new texts was found that had not been seen in the 3.0m corpus. An analysis of the categories these 55 'new words' belong to is shown in Figure 1.

One of the 'new words', unsurprisingly, is ***COVID***, a clear neologism. (***Corona*** was also picked up, but because there was already a single mention of it – as the "Corona Hospital" (in California) – it was marked as outstanding; see § 4.2.) As with every non-English monitor corpus, the expectation was that a good number of 'new words' would be proper names

[2] The second reference should be to Janicijevic/Walker (1997); and the title of Clear's (1988) paper is "Trawling the language: Monitor corpora" rather than the misquoted "The Monitor Corpus". Both of these errors are unfortunately found all over the metalexicographic literature. More upbeat: As with so much in our field, the term 'monitor corpus' was coined by John Sinclair, in 1982, or thus four decades ago already (Clear 1988, p. 383).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

and (given that English is the only official language of Uganda) also plain English words. This is borne out; these categories make up 16% and 4% respectively, so 20% in all. Proper names and plain English words are not normally items that warrant inclusion in a Lusoga dictionary.

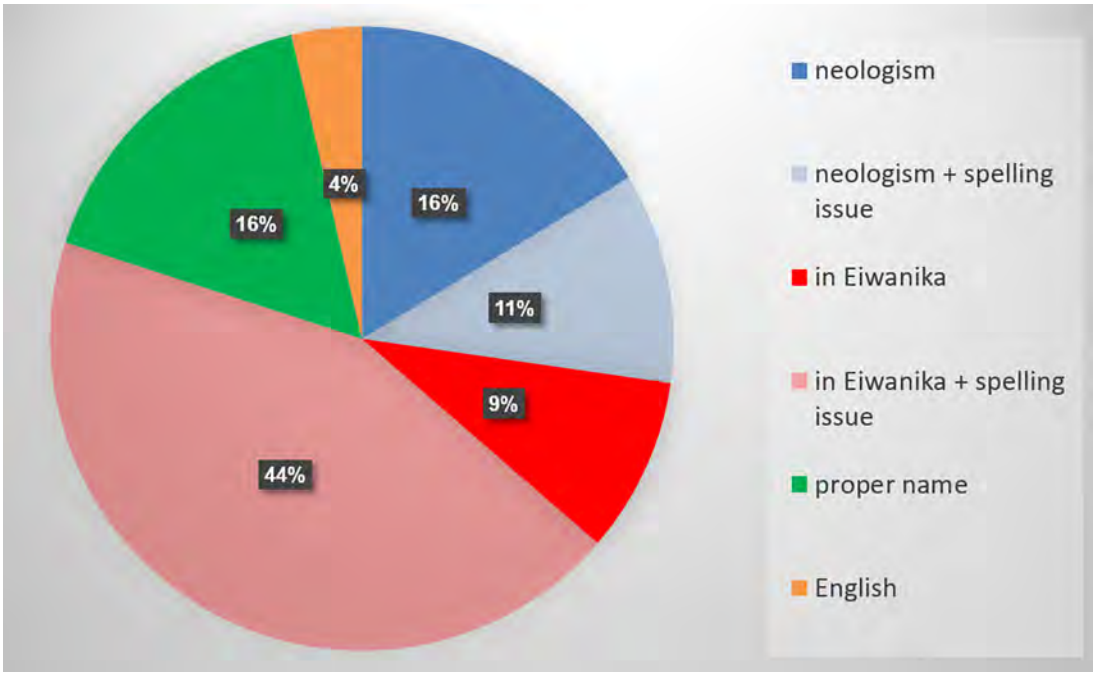


Fig. 1: Analysis of the 'new words' that entered Lusoga between September 2019 and January 2022

In order to analyse the data, it is important to note that no common (standard) orthography has yet been adopted by all those who write in Lusoga, so a tool like WST will also pick up (and give too much weight to) spelling variations – see 'spelling issue' in Figure 1. Had a dictionary been compiled based on the 3.0m corpus, the remaining 80% of the keywords from Figure 1, would all be candidates for inclusion in an update to the dictionary. While such a corpus-based dictionary does not exist, a *non-corpus-based* monolingual dictionary has been compiled, namely the *Eiwanika ly'Olusoga* (Nabirye 2009), online since 2012 at https://menhapublishers.com/dictionary/.

The quest for neologisms may then be rephrased as a quest for candidates to update that dictionary. Astonishingly, as many as 53% of the 'new words' from Figure 1 were already included in the *Eiwanika*, so they are not new words at all; just 27% are. The latter include new loanwords for ***omusaseredooti*** 'priest' (< Latin *sacerdos* 'priest'), ***mwepisikoopi*** 'bishop' (< Latin *episcopus* 'bishop'), and ***ukarisitia*** 'Eucharist' (< Greek *Eucharist* 'gratitude'), but also concepts that can only be 'derived', using language-internal processes, from other words already in the *Eiwanika*, and which are thus debatable neologisms, such as ***obukurisitu*** 'Christianity' (***Omukristo*** 'Christian' is in the *Eiwanika*), ***omuyumo*** 'entertainer' (***ekinhumo*** 'party' is in), or the reduplicated form ***mutoto*** 'youngish' (***-to*** 'young' is in). Conversely, others are clearly true neologisms: ***akanhomero*** 'a small pejorative place' (< ***okunhooma*** 'despise') or ***ekizezengere*** 'shadow' (the personification of ***ekinzenze*** 'a shadow').

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Regarding the first three religious terms here (for 'priest', 'bishop', and 'Eucharist') one may wonder why we label them neologisms, as surely terms for those concepts were already in the language. Suffice it to say that competing religious groups devised their own terms in Lusoga, and that with the recent publication and now inclusion in the Lusoga corpus of Roman Catholic material, these 'new' terms (for old concepts) have now also officially entered the Lusoga language.[3]

4.2 New meanings

In addition to the 55 'new words', WST also lists 1,251 'outstanding words': 815 'positive keywords' (= words that are relatively more frequent in the new 0.5m material compared to the monitor corpus of 3.0m), and 436 'negative keywords' (= words that are relatively less frequent in the new material compared to the monitor corpus, and may thus be 'disappearing from the language'). Of the positive keywords, 466 occur in at least two of the new 0.5m corpus files, while 349 occur in just one of the new files. For the purposes of the present paper, we will only look at the top 100 positive keywords that occur in at least two new texts. An analysis of the categories these 'top 100' belong to is shown in Figure 2.

In Figure 2, the proportion of proper names has slightly grown compared to Figure 1 (to 20%), while that of English has gone down (to 2%). A notable proper name that is now far more outstanding is that of ***Gabula***, the title of the current Busoga King. In terms of candidate new meanings, as many as 61% turn out to have been properly covered in the *Eiwanika*, with their various meanings; yet 17% have not. Some of these 17% indicate that a number of function words which are the result of grammatical constructions had better been lemmatised in the *Eiwanika*, such as the connectives (construction = pronominal prefix + ***-a***), and that some combinations also warrant lemma-sign status, such as ***-liwo*** 'be present' (< ***-li*** 'to be' + ***wo*** (locative)), or ***me ni*** 'and then' (< ***me*** 'and then' + ***ni*** (focus)). These, of course, are neither new words nor new uses; yet the software has (correctly!) picked them out as candidate entries. So here the use of a monitor corpus for Lusoga has not detected new meanings, but forces lexicographers to face the facts; and the fact is that more grammar needs to be entered into the central lemma-sign list of a dictionary.

[3] The work concerned is the Roman missal (Gonza 2018); which despite being dated 2018 only became available in late 2019, whereupon it was scanned, OCRed, and heavily processed (by the first author of this paper, to take out all the English parts) before it was added to the corpus. The Protestant Bible (BSU 2014) was already in the corpus.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

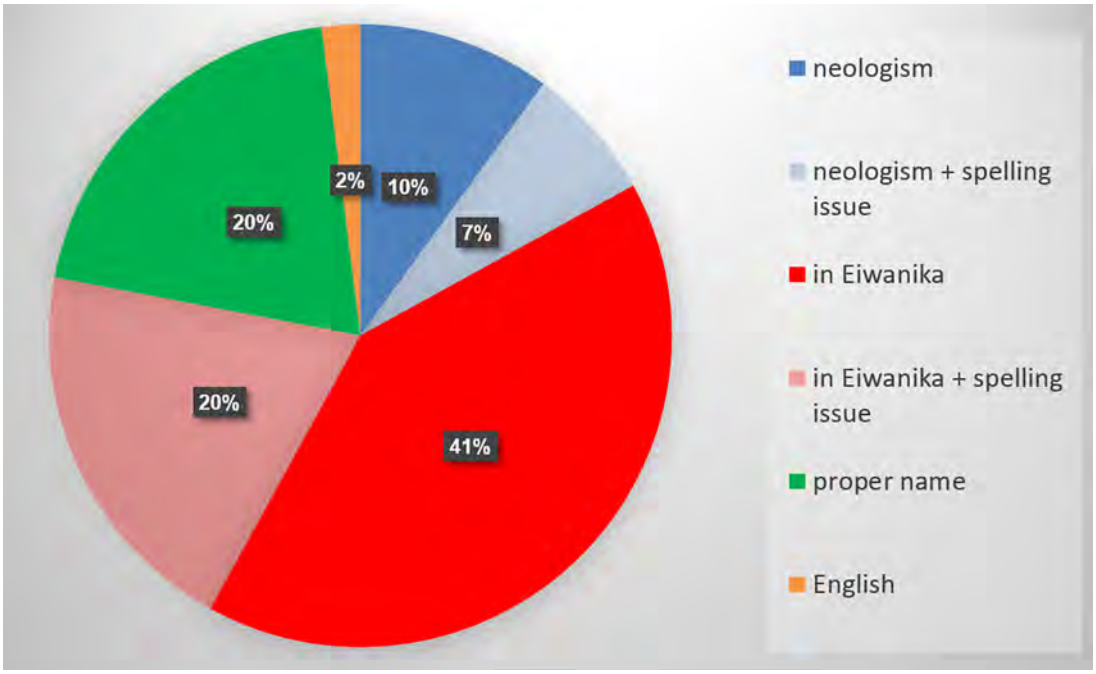


Fig. 2: Analysis of the top 100 'outstanding words' (in at least two corpus files) when comparing Lusoga between September 2019 and January 2022

Full words not lemmatised in the *Eiwanika* include ***lebe*** 'so and so', as well as the interjection ***eee***. The specific but non-descript meaning 'so and so' may be considered a near-neologism; it was hardly there before but now entered the language 'in force'. Similarly for the unspecific interjection *eee*, while not lemmatised in the *Eiwanika*, it *was* used once in a single example (under the lemma ***(a)keewuunia***).

An interesting language change is ***eisakaramentu*** 'sacrament': ***saakalamentu*** was lemmatised in the *Eiwanika*, but the monitor corpus now indicates that the form with a noun class prefix has become far more acceptable than it used to be.

The remainder are all clear cases of neologisms, as these are words that acquired new and very specific meanings. These include: ***ebyeghongo*** 'things used to pray; gifts' (deverbative < ***okuwonga*** 'to give offerings in church'), ***amaingira*** 'the process of entering' (deverbative < ***okwingila*** 'to enter'), ***ekitaloodheka*** 'that which is difficult to relay' (deverbative < cl. 7 noun prefix + negative marker ***-ta*** + ***okuloodha*** 'to relay' + stative extension), ***olugololiro*** 'in a straight manner' (deverbative < ***okugolola*** 'to make straight'), and ***kituufu*** 'it is true' (adjective < ***obutuufu*** 'truthfulness').

5. Discussion and conclusion

As Kilgarriff (2013, p. 82) correctly pointed out: "The nature of the task is that the automatic process creates a list of candidates, and a lexicographer then goes through them to sort the wheat from the chaff. There is always far more chaff than wheat." In terms of 'new words', adding half a million Lusoga tokens to a corpus of 3 million tokens, revealed just 55 items, so having to sort the wheat (which turned out to be 27%) from the chaff manually for such a small amount is more than doable. In terms of 'new meanings', we presented an analysis of the top 100 outstanding words only, where we saw that the wheat was less

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

forthcoming (17%). The full details of going from raw data to analysis may be found in Addenda 1 and 2.[4]

While Kilgarriff does not give us an indication of an acceptable ratio of wheat to chaff, apart from informing us that it is inherently low, we feel that the exercise for Lusoga was worthwhile, as we did pinpoint enough useful material to update the *Eiwanika*. As a result, we are confident that the dawn of monitor corpora for the Bantu languages has arrived.

However, upon also considering recall and precision when going down the list of potential new meanings [to be presented during the actual talk only, as space constraints do not allow for a full description here], we are dealing with a case of diminishing returns: The recall does indeed go up, but at an increasingly punishing precision. Another bottleneck, especially with hopes of automating the process in future, revolves around the various spellings used among the Basoga community; but this is a language-specific problem, not a Bantu-wide one.

References

BSU. 2014. Baibuli. Ekibono kya Katonda. Omuli n'ebitabo ebyetebwa deuterokanoniko/apokurifa [Bible. The word of God, which also has the books known as Deuteronomy/Apocrypha]. Kampala.

Clear, J. (1988): Trawling the language: monitor corpora. In: Snell-Hornby, M. (ed.): ZuriLEX '86 Proceedings: Papers read at the EURALEX International Congress. Tübingen, pp. 383–389.

De Pauw, G./de Schryver, G.-M./Pretorius, L./Levin, L. (2011): Introduction to the special issue on African Language Technology. In: Language Resources and Evaluation 45 (3), pp. 263–269.

de Schryver, G.-M. (2007): Oxford bilingual school dictionary: Northern Sotho and English / Pukuntšu ya Polelopedi ya Sekolo: Sesotho sa Leboa le Seisimane. E gatišitšwe ke Oxford. Cape Town.

de Schryver, G.-M. (2010): Oxford bilingual school dictionary: Zulu and English / Isichazamazwi Sesikole Esinezilimi Ezimbili: IsiZulu NesiNgisi, Esishicilelwe abakwa-Oxford. Cape Town.

de Schryver, G.-M. (2020): Linguistics terminology and neologisms in Swahili: rules vs. practice. In: Dictionaries: Journal of The Dictionary Society of North America 41 (1), pp. 83–104 + 14 pages of supplementary material online.

de Schryver, G.-M./Nabirye, M. (2018): Corpus-driven Bantu lexicography, Part 1: Organic corpus building for Lusoga. In: Lexikos 28, pp. 32–78.

de Schryver, G.-M./Prinsloo, D. J. (2000): The compilation of electronic corpora, with special reference to the African languages. In: Southern African Linguistics and Applied Language Studies 18 (1–4), pp. 89–106.

de Schryver, G.-M./Reynolds, M. (2014): Oxford bilingual school dictionary: IsiXhosa and English / Oxford isiXhosa-isiNgesi English-isiXhosa Isichazi-magama Sesikolo. Cape Town.

Diki-Kidiri, M. (ed.) (2008): Le vocabulaire scientifique dans les langues africaines. Pour une approche culturelle de la terminologie. (= Collection Dictionnaires et Langues). Paris.

[4] Note that it has not been the purpose of this paper to also analyse the various linguistic strategies used to create neologisms in Lusoga, even though some of the evidence for this may be deduced from the Addenda. For a recent Bantu example in this domain, see de Schryver (2020) who deals with term creation processes in Swahili. In Africa more generally, many scholars have worked on this as well; see for instance the edited collection by Diki-Kidiri (2008) for case studies in West and Central Africa.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Dom, S./de Schryver, G.-M./Bostoen, K. (2020): Kisikongo (Bantu, H16a) present-future isomorphism: a diachronic conspiracy between semantics and phonology. In: Journal of Historical Linguistics 10 (2), pp. 251–288.

Gonza, R. K. (2018): Misaale mu Lusoga. Jinja.

Hanks, P. (2002): Mapping meaning onto use. In: Corréard, M.-H. (ed.): Lexicography and natural language processing. A festschrift in honour of B.T.S. Atkins. Euralex, pp. 156–198.

Hanks, P. (2003): Lexicography. In: Mitkov, R. (ed.): The Oxford handbook of computational linguistics. Oxford, pp. 48–69.

Janicijevic, T./Walker, D. (1997): NeoloSearch: Automatic detection of neologisms in French Internet documents. In: Proceedings of the 1997 Joint International Conference of the Association for Computers and the Humanities and the Association for Literary & Linguistic Computting. Kingston, Ontario, pp. 93–94.

Kawalya, Deo. 2017. A corpus-driven study of the expression of modality in Luganda (Bantu, JE15). PhD dissertation. Ghent.

Kawalya, D./Bostoen, K./de Schryver, G.-M. (2021): A diachronic corpus-driven study of the expression of possibility in Luganda (Bantu, JE15). In: International Journal of Corpus Linguistics 26 (3), pp. 336–369.

Kilgarriff, A. (2013): Using corpora as data sources for dictionaries. In: Jackson, H. (ed.): The Bloomsbury companion to lexicography. London, pp. 77–96.

Kosem, I. (2022): Trendi – a monitor corpus of Slovene. In: Klosa-Kückelhaus, A./Engelberg, St./Möhrs, Ch./Storjohann, P. (eds): Proceedings of the XX EURALEX International Congress: Dictionaries and Society. Mannheim.

Kosem, I./Krek, S./Gantar, P./Holdt, Š. A. /Čibej, J. (2021): Language monitor: tracking the use of words in contemporary Slovene. In: Kosem, I./Cukr, M./Jakubíček, M. /Kallas, J./Krek, S./Tiberius, C. (eds): Electronic Lexicography in the 21st Century (eLex 2021): Post-Editing Lexicography. Proceedings. Brno, pp. 514–528.

Mberamihigo, F. (2014): L'expression de la modalité en kirundi: exploitation d'un corpus électronique. PhD dissertation. Brussels/Ghent.

McEnery, T./Hardie, A. (2012): Corpus linguistics: method, theory and practice (= Cambridge Textbooks in Linguistics). Cambridge.

McEnery, T./Wilson, A. (2001): Corpus linguistics: an introduction. 2nd Edition. (= Edinburgh Textbooks in Empirical Linguistics). Edinburgh.

Misago, M.-J. (2018): Les verbes de mouvement et l'expression du lieu en kirundi (bantou, JD62): une étude linguistique basée sur un corpus. PhD dissertation. Ghent.

Misago, M.-J./Nshimirimana, E./Tuyubahe, P. (2021): Usages grammaticaux du verbe -guma 'rester' en kirundi (JD62): Une étude linguistique basée sur un corpus. In: Language in Africa 2 (1), pp. 3–40.

Nabirye, M. (2009): Eiwanika ly'Olusoga. Eiwanika ly'aboogezi b'Olusoga n'abo abenda okwega Olusoga [A dictionary of Lusoga. For speakers of Lusoga, and for those who would like to learn Lusoga] (= Linguistics Series 1). Kampala.

Nabirye, M. (2016): A corpus-based grammar of Lusoga. PhD dissertation. Ghent.

Nabirye, M. (2019): Owayanga: Empayo Dhimala Dhaavaamu Olufumo [Speak regularly: conversations have a tendency to become legends]. (= Linguistics Series 5). Kampala.

Nshemezimana, E. (2016): Morphosyntaxe et structure informationnelle en kirundi: Focus et stratégies de focalisation. PhD dissertation. Ghent.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Prinsloo, D. J./de Schryver, G.-M. (2001): Monitoring the stability of a growing organic corpus, with special reference to Sepedi and Xitsonga. In: Dictionaries: Journal of The Dictionary Society of North America 22, pp. 85–129.

Scott, M. (2019): WordSmith Tools, version 7. http://www.lexically.net/wordsmith/.

UBOS (2016): The National Population and Housing Census 2014 – Main Report. Kampala.

Contact information

Gilles-Maurice de Schryver
BantUGent – UGent Centre for Bantu Studies, Ghent University
& Department of African Languages, University of Pretoria
gillesmaurice.deschryver@UGent.be

Minah Nabirye
BantUGent – UGent Centre for Bantu Studies, Ghent University
minah.nabirye@UGent.be

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Addendum 1: Raw data for 'new words' that entered Lusoga between September 2019 and January 2022

Keyword	Translation	Spelling issue?	Other spelling	Eiwan-ika?	Neolo-gism?	Notes	Freq.	Freq. %	Texts	RC freq.	RC %	BIC	Log-likelihood	Log-ratio	Probability
kurisitu	proper name	✓	Kristu	✓			2,126	0.3933	2	0	0.000	8,100.99	8,116.10	141.01	0.00000000000000000002
jamiili	proper name						142	0.0263	4	0	0.000	526.98	542.09	137.11	0.00000000000000006099
omusaseredooti	priest				✓	< Latin sacerdos 'priest'	137	0.0253	4	0	0.000	507.89	523.00	137.05	0.00000000000000006842
abakurisitu	Christians	✓	Abakristu	✓			90	0.0166	3	0	0.000	328.47	343.58	136.45	0.00000000000000026963
kaliisi	proper name						60	0.0111	2	0	0.000	213.94	229.05	135.86	0.00000000000000107804
musaseredooti	priest				✓	< Latin sacerdos 'priest'	52	0.0096	3	0	0.000	183.40	198.51	135.66	0.00000000000000179678
mbotana	proper name						52	0.0096	2	0	0.000	183.40	198.51	135.66	0.00000000000000179678
oketcho	proper name						50	0.0092	4	0	0.000	175.77	190.88	135.60	0.00000000000000207210
ekiir	abbr. for ekiiribwamu 'response'				✓	ekiiribwamu 'response'	50	0.0092	2	0	0.000	175.77	190.88	135.60	0.00000000000000207210
abakuristu	Christians	✓	Abakristu	✓			48	0.0089	2	0	0.000	168.13	183.24	135.54	0.00000000000000240667
mwepisikoopi	bishop				✓	< Latin episcopus 'bishop'	45	0.0083	2	0	0.000	156.68	171.79	135.45	0.00000000000000305803
omukwampa	the direction of				✓		44	0.0081	2	0	0.000	152.86	167.97	135.42	0.00000000000000332697
ukarisitia	Eucharist				✓	< Greek Eucharist 'gratitude'	35	0.0065	2	0	0.000	118.50	133.61	135.09	0.00000000000000809336
eikukubira	crowd / mob	✓	eikuukuubila	Y			29	0.0054	2	0	0.000	95.60	110.71	134.81	0.00000000000001772373
kategere	proper name						28	0.0052	2	0	0.000	91.78	106.89	134.76	0.00000000000002065687
kukusengera	to praise you	✓	kukusengela	✓			22	0.0041	2	0	0.000	68.88	83.99	134.42	0.00000000000006405400
ainjira	you enter	✓	aingila	✓			20	0.0037	3	0	0.000	61.24	76.35	134.28	0.00000000000010531816
mwoyogwe	his soul	✓	mwoyo gwe	✓			18	0.0033	3	0	0.000	53.61	68.72	134.13	0.00000000000019120947
bukobelana	proper name						18	0.0033	3	0	0.000	53.61	68.72	134.13	0.00000000000019120947
yaacuukira	he changed from	✓	yakyukila	✓			17	0.0031	3	0	0.000	49.79	64.90	134.04	0.00000000000027119964
fridge	English						16	0.0030	2	0	0.000	45.97	61.08	133.96	0.00000000000040281970
kirimungo	proper name						15	0.0028	4	0	0.000	42.15	57.26	133.86	0.00000000000063545448
kuntwiko	on top	✓	ku ntwiko	✓			15	0.0028	2	0	0.000	42.15	57.26	133.86	0.00000000000063545448

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Keyword	Translation	Spelling issue?	Other spelling	Eiwan-ika?	Neolo-gism?	Notes	Freq.	Freq. %	Texts	RC freq.	RC %	BIC	Log-likelihood	Log-ratio	Probability
mwikuukuubira	in a crowd / mob	✓	mw'ikuukuubi-la	✓			15	0.0028	2	0	0.000	42.15	57.26	133.86	0.000000000000063545448
kwigulu	on heaven	✓	ku igulu	✓			14	0.0026	2	0	0.000	38.34	53.45	133.76	0.000000000000108799153
asikaali	guard			✓			14	0.0026	2	0	0.000	38.34	53.45	133.76	0.000000000000108799153
covid	COVID				✓	COVID & COVID-19	13	0.0024	6	0	0.000	34.52	49.63	133.66	0.000000000000209591906
omuyumo	entertainer				✓	**ekinhumo** 'party' is in; but not the entertainer	13	0.0024	5	0	0.000	34.52	49.63	133.66	0.000000000000209591906
omuminsani	missionary			✓			13	0.0024	2	0	0.000	34.52	49.63	133.66	0.000000000000209591906
obukurisitu	Christianity	✓	Obukristu		✓	**Omukristo** 'Christian' is in; but not cl. 14	12	0.0022	3	0	0.000	30.70	45.81	133.54	0.000000000000485604720
bwoisa	as you breathe / as you say	✓	bw'oisa	✓			12	0.0022	2	0	0.000	30.70	45.81	133.54	0.000000000000485604720
leonard	proper name						12	0.0022	2	0	0.000	30.70	45.81	133.54	0.000000000000485604720
okugwana	to deserve			✓			12	0.0022	2	0	0.000	30.70	45.81	133.54	0.000000000000485604720
akanhomero	a small pejorative place	✓	akanhoomelo		✓	**okunhooma** 'despise' is in; but not the deverbative	12	0.0022	2	0	0.000	30.70	45.81	133.54	0.000000000000485604720
omubirigwe	his body	✓	omubili gwe	✓			11	0.0020	3	0	0.000	26.88	41.99	133.42	0.000000000001566624434
acuuka	he changes	✓	akyuka	✓			11	0.0020	3	0	0.000	26.88	41.99	133.42	0.000000000001566624434
buliribwe	his bed	✓	bulili bwe	✓			11	0.0020	2	0	0.000	26.88	41.99	133.42	0.000000000001566624434
abamwaguleku	so that he sprinkles on you				✓	**okumwaga** 'to sprinkle' is in; but not the reversive + subjunctive + loc. form	10	0.0018	2	0	0.000	23.07	38.18	133.28	0.000000000010989710877
filename	English						9	0.0017	9	0	0.000	19.25	34.36	133.13	0.000000001660683479621
bakurisitu	Christians	✓	Bakristu	✓			9	0.0017	3	0	0.000	19.25	34.36	133.13	0.000000001660683479621
bwanaaba	if he will be	✓	bw'anaaba	✓			9	0.0017	3	0	0.000	19.25	34.36	133.13	0.000000001660683479621
simwogerere	proper name						9	0.0017	3	0	0.000	19.25	34.36	133.13	0.000000001660683479621
mwikonero	home	✓	mw'ikonelo		✓	new meaning (was: chair; water container > home > chair; whomever owns the water container owns the home, thus chair > cf. one is not supposed to sit in the parent's chair)	9	0.0017	2	0	0.000	19.25	34.36	133.13	0.000000001660683479621
yaacuuka	he is changing	✓	yaakyuka	✓			9	0.0017	2	0	0.000	19.25	34.36	133.13	0.000000001660683479621
endhuuye	his house	✓	endhu ye	✓			9	0.0017	2	0	0.000	19.25	34.36	133.13	0.000000001660683479621
ebiraalagho	those which remain; those which are maintained	✓	ebilaalawo		✓	**okulaala** 'be calm' is in; -wo = loc.	9	0.0017	2	0	0.000	19.25	34.36	133.13	0.000000001660683479621

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Keyword	Translation	Spelling issue?	Other spelling	Eiwan-ika?	Neolo-gism?	Notes	Freq.	Freq. %	Texts	RC freq.	RC %	BIC	Log-likelihood	Log-ratio	Probability
ekizezengere	shadow	✓	ekinzenzegele		✓	**ekinzenze** 'a shadow' is in Eiwanika; but new personification of 'shadow' is not	9	0.0017	2	0	0.000	19.25	34.36	133.13	0.000000001660683479621
ekiyumu	party	✓	ekinhumo	✓			8	0.0015	8	0	0.000	15.43	30.54	132.96	0.000000029774795962112
bainjira	they enter	✓	baingila	✓			8	0.0015	4	0	0.000	15.43	30.54	132.96	0.000000029774795962112
kyamwidhira	it came to him	✓	kyamwidhila	✓			8	0.0015	3	0	0.000	15.43	30.54	132.96	0.000000029774795962112
twamubaamu	we were in it			✓		-ba	8	0.0015	3	0	0.000	15.43	30.54	132.96	0.000000029774795962112
okucaala	to visit	✓	okukyala	✓			8	0.0015	3	0	0.000	15.43	30.54	132.96	0.000000029774795962112
mutoto	youngish	✓	mutooto		✓	-to 'young' is in; but not the reduplicated form	8	0.0015	2	0	0.000	15.43	30.54	132.96	0.000000029774795962112
kabbuli	grave (Islam)			✓			8	0.0015	2	0	0.000	15.43	30.54	132.96	0.000000029774795962112
bweyinza	as it can	✓	bw'eyinza	✓			8	0.0015	2	0	0.000	15.43	30.54	132.96	0.000000029774795962112

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Addendum 2: Raw data for the top 100 'outstanding words' (in at least two corpus files) when comparing Lusoga between September 2019 and January 2022

Keyword	Translation	Spelling issue?	Other spelling	*Eiwanika*?	Neologism?	Notes	Freq.	Freq. %	Texts	RC freq.	RC %	BIC	Log-likelihood	Log-ratio	Probability
amiina	Amen			✓			1,818	0.3363	9	538	0.017	4,566.20	4,581.31	4.28	0.000000000000000000009
ebyanda	century			✓			1,989	0.3680	6	1,506	0.048	3,283.21	3,298.31	2.92	0.000000000000000000023
ighe	you	✓	iwe	✓			1,360	0.2516	17	523	0.017	3,119.60	3,134.71	3.90	0.000000000000000000027
tubita	we pass			✓			1,447	0.2677	9	892	0.029	2,685.58	2,700.69	3.22	0.000000000000000000042
tukweghembye	we have beseached you	✓	tukwewembye	✓			1,035	0.1915	2	321	0.010	2,554.86	2,569.97	4.21	0.000000000000000000049
tukweghemba	we beseach you	✓	tukwewemba	✓			1,394	0.2579	3	925	0.030	2,484.12	2,499.23	3.11	0.000000000000000000053
mutukuvu	he is holy			✓			1,222	0.2261	4	704	0.023	2,346.92	2,362.03	3.32	0.000000000000000000063
alamula	he passes judgement			✓			700	0.1295	3	62	0.002	2,247.16	2,262.27	6.02	0.000000000000000000072
aligho	he is present	✓	aliwo		✓	-li 'to be' is in; but not the combined form	715	0.1323	9	95	0.003	2,159.33	2,174.44	5.43	0.000000000000000000082
waife	ours			✓			1,926	0.3563	26	2,458	0.079	2,113.59	2,128.70	2.17	0.000000000000000000087
nabirye	proper name						2,827	0.5230	23	5,068	0.163	2,103.82	2,118.93	1.68	0.000000000000000000088
okusembera	to come close; to receive communion	✓	okusembela	✓			831	0.1537	3	493	0.016	1,567.30	1,582.40	3.28	0.000000000000000000215
lw	connective 'of'				✓	-a 'of' not in	1,438	0.2660	33	2,089	0.067	1,376.38	1,391.49	1.98	0.000000000000000000320
olwembo	song	✓	olwemba	✓			920	0.1702	9	825	0.027	1,347.91	1,363.02	2.68	0.000000000000000000341
omutuukirivu	the one who is holy	✓	omutuukilivu	✓			575	0.1064	3	216	0.007	1,321.79	1,336.90	3.93	0.000000000000000000362
nantambi	proper name						442	0.0818	5	96	0.003	1,198.39	1,213.50	4.73	0.000000000000000000488
mwoyo	soul			✓			1,097	0.2029	15	1,499	0.048	1,117.53	1,132.64	2.07	0.000000000000000000603
okughonga	to pray	✓	okuwonga	✓			575	0.1064	3	323	0.010	1,110.43	1,125.54	3.35	0.000000000000000000615
kaluuba	proper name						345	0.0638	2	30	0.001	1,102.49	1,117.60	6.05	0.000000000000000000629
wo	locative 16			✓			1,582	0.2927	22	3,000	0.097	1,081.22	1,096.33	1.60	0.000000000000000000667
yezu	proper name		Yesu / Yezu	✓			727	0.1345	3	645	0.021	1,070.17	1,085.28	2.69	0.000000000000000000689

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Keyword	Translation	Spelling issue?	Other spelling	*Eiwanika*?	Neologism?	Notes	Freq.	Freq. %	Texts	RC freq.	RC %	BIC	Log-likelihood	Log-ratio	Probability
ebyeghongo	things used to pray; gifts	✓	ebyewongo		✓	okuwonga 'to give offerings in church' is in; but not the deverbative	630	0.1165	3	480	0.015	1,025.55	1,040.66	2.91	0.00000000000000000784
abailamu	those who respond			✓			275	0.0509	5	2	0.000	1,011.65	1,026.76	9.63	0.00000000000000000818
ndhote	proper name						284	0.0525	7	16	0.001	949.28	964.39	6.67	0.00000000000000000993
eighanga	country; tribe	✓	eiwanga	✓			651	0.1204	11	598	0.019	932.81	947.92	2.64	0.00000000000000001048
kadaaga	proper name						303	0.0561	5	46	0.001	884.29	899.40	5.24	0.00000000000000001234
kasaata	proper name						248	0.0459	9	9	0.000	856.52	871.62	7.31	0.00000000000000001361
mata	proper name						355	0.0657	14	118	0.004	846.57	861.68	4.11	0.00000000000000001411
omughanzi	the one who is able	✓	omuwanzi	✓			443	0.0820	3	265	0.009	824.85	839.96	3.26	0.00000000000000001528
lebe	so and so				✓	MISSED	334	0.0618	7	104	0.003	813.18	828.29	4.21	0.00000000000000001596
aghalala	together; collectively	✓	awalala	✓			792	0.1465	13	1,076	0.035	807.48	822.59	2.08	0.00000000000000001631
baagalana	proper name						232	0.0429	2	8	0.000	802.98	818.09	7.38	0.00000000000000001660
esaala	prayer			✓			500	0.0925	10	407	0.013	776.47	791.58	2.82	0.00000000000000001840
kelezia	chapel	✓	keleziya	✓			403	0.0746	4	225	0.007	776.13	791.24	3.36	0.00000000000000001842
yafeesi	office	✓	yaafeesi	✓			302	0.0559	3	87	0.003	752.21	767.32	4.32	0.00000000000000002029
zabbuli	psalms			✓			253	0.0468	2	35	0.001	748.87	763.98	5.38	0.00000000000000002057
amaingira	the process of entering	✓	amaingila		✓	okwingila 'to enter' is in; but not the deverbative	399	0.0738	2	238	0.008	742.55	757.65	3.27	0.00000000000000002111
okweyanza	to say thank you			✓			418	0.0773	2	278	0.009	733.34	748.45	3.11	0.00000000000000002194
okulokolwa	to become saved			✓			230	0.0425	2	23	0.001	716.16	731.27	5.84	0.00000000000000002360
eee	(interjection)				✓	(used in 1 eg of the *Eiwanika*)	452	0.0836	26	365	0.012	704.25	719.36	2.83	0.00000000000000002485
obutikitiki	seconds			✓			192	0.0355	3	9	0.000	647.25	662.36	6.94	0.00000000000000003225
rose	proper name						246	0.0455	2	59	0.002	643.33	658.43	4.58	0.00000000000000003286
ekisa	mercy			✓			481	0.0890	10	485	0.016	637.65	652.76	2.51	0.00000000000000003377
fuuti	foot			✓			207	0.0383	4	25	0.001	624.55	639.66	5.57	0.00000000000000003601
mutalya	proper name						197	0.0364	2	20	0.001	609.90	625.01	5.82	0.00000000000000003875
mulala	one (cl. 1/3)			✓			973	0.1800	33	1,961	0.063	600.66	615.76	1.51	0.00000000000000004063

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Keyword	Translation	Spelling issue?	Other spelling	*Eiwanika*?	Neologism?	Notes	Freq.	Freq. %	Texts	RC freq.	RC %	BIC	Log-likelihood	Log-ratio	Probability
amaghango	laws	✓	amawango	✓			186	0.0344	2	15	0.000	593.06	608.17	6.15	0.00000000000000004226
ganha	allow			✓			347	0.0642	4	241	0.008	591.00	606.11	3.05	0.00000000000000004272
muli	you are			✓		< SM + -li	814	0.1506	29	1,467	0.047	590.65	605.76	1.67	0.00000000000000004280
tughange	so that we allow	✓	tuwange	✓			275	0.0509	4	135	0.004	558.44	573.55	3.55	0.00000000000000005093
siraji	proper name						186	0.0344	7	25	0.001	549.42	564.52	5.42	0.00000000000000005358
meeni	and then	✓	me ni		✓	< me 'and then' + ni (focus); me not in	238	0.0440	6	86	0.003	546.09	561.20	3.99	0.00000000000000005460
byaruhanga	proper name						159	0.0294	2	6	0.000	542.26	557.37	7.25	0.00000000000000005581
dhaakaba	proper name						177	0.0327	2	19	0.001	541.92	557.03	5.74	0.00000000000000005591
kawala	proper name						161	0.0298	3	8	0.000	537.66	552.77	6.85	0.00000000000000005730
gavumenti	government			✓			719	0.1330	23	1,332	0.043	499.97	515.08	1.63	0.00000000000000007185
nh	and	✓	n'		✓	ni 'and'/(focus) is in	346	0.0640	2	318	0.010	488.51	503.61	2.64	0.00000000000000007724
m	proper name						661	0.1223	12	1,182	0.038	482.02	497.13	1.68	0.00000000000000008053
omwembi	singer			✓			193	0.0357	18	69	0.002	441.72	456.82	4.01	0.00000000000000010583
diguli	degree			✓			228	0.0422	9	123	0.004	440.08	455.18	3.41	0.00000000000000010707
mw	connective 'of'				✓	-a 'of' not in	583	0.1079	33	1,028	0.033	431.67	446.78	1.70	0.00000000000000011374
omukulembeze	leader			✓			292	0.0540	13	278	0.009	398.99	414.10	2.59	0.00000000000000014567
gabula	proper name					title of the current Busoga King	255	0.0472	5	210	0.007	385.50	400.61	2.80	0.00000000000000016236
pulezidenti	President			✓			193	0.0357	5	100	0.003	377.62	392.73	3.47	0.00000000000000017329
abatuukirivu	holy people	✓	abatuukilivu	✓			241	0.0446	4	189	0.006	375.78	390.89	2.87	0.00000000000000017600
ekigwaine	what is deserving			✓			145	0.0268	3	35	0.001	372.33	387.44	4.57	0.00000000000000018119
mbeeni	preferably			✓			192	0.0355	5	106	0.003	363.94	379.05	3.38	0.00000000000000019474
ziraba	proper name						117	0.0216	2	12	0.000	355.55	370.66	5.81	0.00000000000000020966
tughe	let us give; give us	✓	tuwe	✓		< SM + -wa + subjunctive	275	0.0509	3	289	0.009	345.95	361.06	2.45	0.00000000000000022867
eisakaramentu	religious teaching	✓	eisaakalamen-tu		✓	saakalamentu 'sacrament' is in; now with prefix	101	0.0187	2	3	0.000	344.24	359.35	7.60	0.00000000000000023230
nakaziba	proper name						111	0.0205	3	10	0.000	342.83	357.94	5.99	0.00000000000000023533

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Keyword	Translation	Spelling issue?	Other spelling	*Eiwan-ika*?	Neolo-gism?	Notes	Freq.	Freq. %	Texts	RC freq.	RC %	BIC	Log-likelihood	Log-ratio	Probability
bagiira	they go / they follow	✓	bagiila	✓			101	0.0187	4	5	0.000	331.77	346.88	6.86	0.00000000000000000026120
omwepisikoopi	bishop				✓	< Latin episcopus 'bishop'	107	0.0198	3	10	0.000	328.27	343.38	5.94	0.00000000000000000027017
yoanne	proper name						234	0.0433	2	225	0.007	314.28	329.39	2.58	0.00000000000000000031040
omusengo	offerings			✓			272	0.0503	4	316	0.010	312.84	327.95	2.31	0.00000000000000000031498
meemale	before; after			✓			280	0.0518	22	337	0.011	311.89	327.00	2.25	0.00000000000000000031803
abe	so that (s)he is			✓		< SM + -ba + subjunctive	487	0.0901	15	956	0.031	305.68	320.79	1.55	0.00000000000000000033916
byona	also (pl.)			✓			496	0.0918	21	995	0.032	301.05	316.15	1.52	0.00000000000000000035615
baghange	so that they can	✓	okuwange	✓			145	0.0268	4	67	0.002	295.43	310.54	3.64	0.00000000000000000037830
mosese	proper name						94	0.0174	2	7	0.000	295.11	310.22	6.27	0.00000000000000000037958
ekitaloodheka	that which is difficult to relay				✓	okuloodha 'to relay' is in; now in cl. 7 + neg. 'ta' + stat. ext.	124	0.0229	3	40	0.001	288.89	304.00	4.15	0.00000000000000000040645
kulw	because of	✓	ku lw'		✓	-a 'of' not in	118	0.0218	6	33	0.001	287.38	302.49	4.36	0.00000000000000000041332
boss	English						107	0.0198	2	21	0.001	285.84	300.95	4.87	0.00000000000000000042052
aghooni	there then	✓	awo ni	✓		< awo 'there' + ni (focus)	90	0.0166	4	6	0.000	285.51	300.62	6.43	0.00000000000000000042211
lugololiro	in a straight manner				✓	okugolola 'to make straight' is in; deverbative is not	83	0.0154	2	3	0.000	276.68	291.79	7.31	0.00000000000000000046697
kimoimo	sole; alone			✓			118	0.0218	2	44	0.001	259.99	275.10	3.95	0.00000000000000000057081
olugololiro	in a straight manner				✓	okugolola 'to make straight' is in; deverbative is not	90	0.0166	2	12	0.000	258.43	273.54	5.43	0.00000000000000000058205
aka	which belongs to				✓	-a 'of' not in	196	0.0363	16	196	0.006	252.61	267.72	2.52	0.00000000000000000062663
omutukuvu	the one who is holy			✓			314	0.0581	5	523	0.017	243.87	258.98	1.79	0.00000000000000000070249
ebyamagero	miracles	✓	ebyamagelo	✓			112	0.0207	3	46	0.001	236.62	251.73	3.81	0.00000000000000000077509
masaaba	place name			✓			75	0.0139	2	5	0.000	235.40	250.51	6.43	0.00000000000000000078819
kituufu	it is true				✓	only cl. 14 obutuufu 'truthfulness' is in; here used as an adj.	309	0.0572	20	523	0.017	234.64	249.75	1.76	0.00000000000000000079660
siima	appreciate			✓			152	0.0281	2	124	0.004	225.18	240.29	2.82	0.00000000000000000091125
cd	English			✓			69	0.0128	14	3	0.000	224.32	239.43	7.05	0.00000000000000000092276
abagololeraku	those who raise [the hand]			✓			65	0.0120	2	1	0.000	222.99	238.10	8.54	0.00000000000000000094098
omwegha	virgin			✓			180	0.0333	2	195	0.006	215.37	230.48	2.41	0.00000000000000000105466

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Keyword	Translation	Spelling issue?	Other spelling	*Eiwan-ika*?	Neolo-gism?	Notes	Freq.	Freq. %	Texts	RC freq.	RC %	BIC	Log-likelihood	Log-ratio	Probability
okugabana	to divide up			✓			151	0.0279	5	131	0.004	213.87	228.98	2.73	0.000000000000000107927
olukiiko	meeting			✓			350	0.0647	9	696	0.022	210.99	226.10	1.53	0.000000000000000112860
tuwe	give us			✓			124	0.0229	4	80	0.003	210.70	225.81	3.15	0.000000000000000113359
kiidhukizo	commemoration			✓			70	0.0129	2	8	0.000	203.10	218.21	5.65	0.000000000000000127984

Main abbreviations used in the Addenda

adj. = adjective
BIC = Bayesian Information Criterion
cl. = class
freq. = frequency in new Lusoga material (0.5m)
loc. = locative
RC freq. = frequency in reference corpus (3.0m)
SM = subject marker
stat. ext. = stative extension
texts = number of new texts keyword occurs in

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Phraseology & Collocations

Maria Ermakova/Alexander Geyken/
Lothar Lemnitzer/Bernhard Roll

INTEGRATION OF MULTI-WORD EXPRESSIONS INTO THE DIGITAL DICTIONARY OF GERMAN LANGUAGE (DWDS)

Towards a lexicographic representation of phraseological variation

Abstract One central goal of the project 'Zentrum für digitale Lexikographie der deutschen Sprache' (Center for digital lexicography for the German Language, www.zdl.org) is to provide a corpus-based lexicographic component of common German multi-word expressions (MWE), including idioms, for DWDS (www.dwds.de), a general language dictionary of contemporary German. As a central challenge of this task, we have identified an adequate lexicographic representation of such common properties of MWE as variation and modification. To document the variation, we have developed a special entry-clustering model, which we call *hub-node entry*. This model comprises a core hub entry headed by a short nuclear form of the MWE and several node entries, which represent the most common variants in their full lexical forms.

Keywords Multi-word expressions, phraseological variation, dictionary entry structure

1. Introduction

The Center for Digital Lexicography for the German Language (ZDL), a project that is currently funded by the German Federal Ministry of Education and Research for the period from 2019 through 2023, pursues the goal to describe the German language comprehensively, while remaining true to a scientifically pure approach – on a digital platform that is accessible free of charge for everyone. This platform links several resources, dedicated to synchronic and diachronic descriptions of words. Users can query all resources simultaneously through a single search engine. The ZDL brings together digital dictionaries, both legacy dictionaries and dictionaries that are currently under development, extensive corpus resources covering several centuries of German language history, and digital linguistic tools that visualise syntactic and historical data for specific words.

The ZDL encompasses the "Digitales Wörterbuch der deutschen Sprache" (DWDS – Digital Dictionary of the German Language), a long-term project that is funded by the Berlin-Brandenburg academy of Sciences and the Humanities from 2007 through 2024. The DWDS provides a part of the contemporary language component of the project and the ZDL is currently building upon it. The DWDS provides an information system for the historical and present vocabulary of the German language (www.dwds.de). The various information sources are continuously updated. Both the ZDL and DWDS can be consulted free of charge.

ZDL and DWDS pursue two goals. Firstly, they aim at pooling and updating the lexical information from the large dictionaries published until now. Secondly, they strive to provide an information system that connects a reliable and scientifically sound lexicographical description of words with the possibility of researching the various uses of a word in well-documented text corpora.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

One of the main goals of the project is to fully describe the majority of the frequently used multi-word expressions (MWE), focusing on idiomatic expressions (IE).[1,2]

The major source for the acquisition and description of these lexical items are large corpora of contemporary German as well as several existing lexical resources. Special attention is given to the formal variability of the MWE, a variability that can be observed in our corpora. Many MWE turn out to be semi-fixed, i. e. they can be modified in certain ways but are less flexible than freely composed phrases. A challenge of this project is to appropriately describe the range of regular variation for (groups of) MWE.

In this article, we will firstly review the traditional placement of the MWE in general dictionaries of German and shortly outline, how we integrate MWE in DWDS as full-fledged entries (section 2). The next section (section 3) addresses the types of phraseological variation that we deal with in our project and shows, how we treat the typical cases of moderate phraseological variation in DWDS. Further to this, we will direct the readers' attention on our modelling approach for idiomatic expressions with a wider range of internal variation. For the comprehensive record of the full range of (observed) variations, we developed an architecture of hub entries and node entries. In section 4 we will present and discuss this model. In section 5 we will present our conclusion and suggestions for further work.

2. Idiomatic expressions as part of traditional (print) dictionaries and the DWDS

MWE have since long been the object of lexicographical description, particularly in German lexicography. Up until now, there is a range of specialized dictionaries describing (semi-) fixed expressions such as collocations and idiomatic expressions, e. g. "Deutsche Idiomatik" (ed. H. Schemann) (2011) and Duden 11 (2020). Multi-word expressions of all kinds are also part and parcel of large monolingual general dictionaries of contemporary German. The most prominent examples of this species are the six-volume "Wörterbuch der deutschen Gegenwartssprache" (WDG 1962–1977), and, in the wake of this pioneering work, the one-volume Duden Universalwörterbuch (DUW 2018) and the one-volume Wahrig Deutsches Wörterbuch (WDW 2018). A retro-digitized version of the WDG is also part of the lexical stock of the DWDS.

In general language dictionaries, MWE are usually listed within the microstructure of single-word articles and highlighted with a specific typographic marker like an asterisk or bold type for them to be found easily by the users (Burger 2015, p. 185). WDG/DUW on the one hand and WDW on the other hand treat these items differently. While the former integrate these units in the sense part of the entry and subsume these units under the closest or most appropriate sense of the component word, the WDW does with lists of MWE at the end of

1 We consider idiomatic expressions to be a subclass of multi-word expressions. The latter also contains lexical items such as complex nouns, complex adverbs, light verb constructions, etc.

2 For an overview of the essential properties of multi-word expressions, see Hüning/Schlücker (2015). However, idiomatic expressions pose the largest challenges with regard to a description of their variability. They are therefore dealt with in the rest of the article and MWE and IE are used interchangeably.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

an article, i. e. outside the proper sense descriptions of the headword. Even specialized dictionaries for idiomatic expression order the entries in accordance with this principle: MWE are grouped under the headwords, more precisely: under the first sense bearing headword of the MWE.

In contrast to this approach, our project is going to create full-fledged entries for multi-word expressions and therefore make them a part of the macro-structure. Furthermore, we intertwine both the entries for single word expressions and the entries for multi-word expressions through a system of cross-references and embedding (see Fig. 1 and Fig. 2 below for an example).

MWE are lexical units of their own right[3] and are as complex as single words: a) they sometimes unfold to a broad range of formal variants; b) in many cases they are used in more than one sense; c) each of these senses deserves a full lexicographical description, in particular i) a definition, ii) usage notes about style as well as temporal and local restrictions, iii) typical combinations with other words (i. e. typical collocations and internal as well as external modification), iv) a sufficiently rich set of usage examples drawn from corpora. Such a rich and detailed description is not possible if these MWE were part of the microstructure(s) of their component words[4].

In other words, articles for multi-word expressions follow (with few exceptions) the microstructure of articles for single words. Technically speaking, one and the same XML-schema is used for both single word entries and multi-word entries. Of course, the same holds for the presentation of articles on the website.

On the other hand, MWE are embedded into the articles of their meaning bearing component words. Each single word entry has as one of its parts a list of MWE entries of which it is part (see Fig. 2). This list is automatically generated and draws on the "Bestandteile" (= components) section of the MWE. Figure 1 contains a list of the four meaning bearing components of the idiomatic expression *weit/ weitab vom Schuss* (lit. far/far away from shot; engl.: 'far away; in the far distance').

Fig. 1: A screenshot of the article *weit vom Schuss/ weitab vom Schuss* in the DWDS, cf. https://www.dwds.de/wb/weitab%20vom%20Schuss

[3] Cf. Elsen (2017, p. 147).

[4] Cf. https://www.dwds.de/wb/jmdn.%2C%20etw.%20an%20der%20Angel%20 haben or https://www.dwds.de/wb/den%20Dienst%20versagen to get an impression of the complexity of the lexicographical description of multi word expression.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

The figure below shows a screenshot of the article *Schuss*, where, in the form section, all idiomatic expressions containing this word are listed.

Schuss, der

Grammatik	Substantiv (Maskulinum) · Genitiv Singular: **Schusses** · Nominativ Plural: **Schüsse**
Nebenform	**Schuss** · Substantiv · Nominativ Plural: **Schuss** (bei Mengenangabe)
Aussprache	[ʃʊs]
Ungültige Schreibung	Schuß
Rechtschreibregeln	§ 2, § 25 (E1)
Wortbildung	mit ›Schuss‹ als Erstglied: ↗ Schussbahn ... 35 weitere · mit ›Schuss‹ als Letztglied: ↗ Bauchschuss ... 57 weitere · mit ›Schuss‹ als Binnenglied: ↗ Bolzenschussapparat · ↗ Selbstschussanlage
Mehrwortausdrücke	↗ Schuss in den Ofen · ↗ Schuss ins Blaue · ↗ den Schuss nicht gehört haben · ↗ einen Schuss haben · ↗ fern vom Schuss · ↗ fernab vom Schuss · ↗ im Schuss · ↗ in Schuss · ↗ in Schuss kommen · ↗ nur einen Schuss haben · ↗ sich, etw. im Schuss halten · ↗ sich, etw. in Schuss bringen · ↗ sich, etw. in Schuss halten · ↗ weit vom Schuss · ↗ weitab vom Schuss ... weniger

Fig. 2: A screenshot of the article *Schuss* in the DWDS. Clicking on one of the listed MWE leads the user to the resp. article, cf. https://www.dwds.de/wb/Schuss.

In the following section we will present several examples of MWE entries in DWDS with the focus on the documentation of their phraseological variation.

3. Representation of the phraseological variation in ZDL project: Flexibility of the MWE as a challenge for lexicography

One of the features of MWE is their lexical and semantic stability (Fleischer 1982, p. 41). Variation in the lexical components of MWE is nevertheless a common phenomenon within this segment of the lexicon and cannot be ignored if the goal is to record the full range of recurrent patterns of variation that can be observed in actual language use. There is often more to the picture than meets the lexicographers' intuition (cf. Stumpf 2019, p. 116).

Phraseological variation is quite heterogeneous and comprises alteration of different components within MWE[5]. For example, the expression *am falschen Ende sparen* ('to make savings on the wrong end') can also be used with other nouns: *an der falschen Stelle sparen*, *am falschen Ort sparen*, *am falschen Platz sparen* ('to make savings in the wrong place'). In the expression *hartes Brot* ('hard bread'), which means 'a tough way', the noun can also be used with another adjective, i. e. *schweres Brot* ('heavy bread'). The expression *auf Achse sein* ('to be on the axis') meaning 'to be on journeys' can also be used with a definite article *auf der Achse sein.*

In traditional lexicography, variation is recorded on the basis of the lexicographers' intuition and not cross-checked with evidence from the corpora: many of these descriptions therefore draw an incomplete picture. Variable components are typically put into brackets or are separated from each other by a slash or comma, this leads to very complex forms of headwords that are hard to decode, e. g. from "Deutsche Idiomatik" by Hans Schemann (2011): *die Hand darauf/dadrauf/auf das/ein Versprechen… geben* ('to give one's hand on it / sth. / on a promise …').

[5] For examples cf. Nunberg/Sag/Wasow (1984), Staffeldt (2009, p. 204).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

In the ZDL project, we apply another method for MWE documentation, which avoids brackets and other punctuation symbols for highlighting variation. According to the complexity of the variation(s) on the part of the headword, we treat these either as a single entry or we create separate entries.

For example, in the case of moderate variation such as an alteration of an adjective, two (or more) separate entries are created, one for each variant:

Fig. 3: phraseological variants represented in two dictionary entries: screenshot of the article *hartes Brot* (https://www.dwds.de/wb/hartes%20Brot) and the article *schweres Brot* (https://www.dwds.de/wb/schweres%20Brot) – moderate variation of the head form.

Further cases of moderate variation of the headword, such as the alteration of prepositions or the extension by one word with no alteration in meaning, are dealt within one single entry, e.g. (see Fig. 4): *keinen Dunst haben – keinen blassen Dunst haben* 'to have no (pale) haze' meaning 'to have not the faintest idea'.

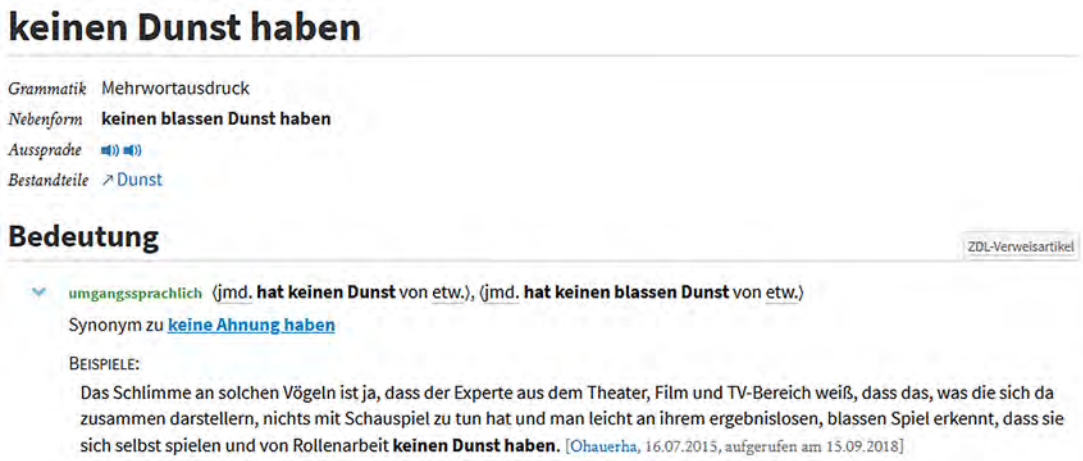


Fig. 4: phraseological variants represented in one dictionary entry: screenshot of the article *keinen Dunst haben* (https://www.dwds.de/wb/keinen%20Dunst%20haben) with the moderate extension of an otherwise stable headword.

More complex are the cases of variation that affect the integrity of the headword itself, with (example 3) or without (example 1) change in meaning. Frequently, the variation is limited to a small and arbitrary set of lexical units (example 2):

(1) *sich verkaufen, laufen, gehen, weggehen … wie geschnitten Brot*
refl. pronoun to sell, to run, to go, to go away … as cut bread
'sell like hot cakes'

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

(2) *sich zum Deppen, Narren, Horst, Affen … machen*
refl. pronoun to the moron, fool, Horst, monkey … make
'to make a fool of oneself'

(3) *am Drücker sitzen, sein, bleiben*
at pusher sit, be, stay
'to be/stay in power (of sth.)'

In short, there are many types of variation. A flexible way to document the full range is necessary. In the following section, we will introduce one of the dictionary entry models that we have been developing for common variation types as the high alteration of the verb in verbal MWE (1, 3). Its function is to capture both the most frequent forms of the MWE as well as other less common variants and modifications, as obtained from our corpus.

4. Hub-node model for groups of MWE entries

As mentioned above, some MWE families are too complex to be described in one or two dictionary entries. In particular, it concerns the MWE with high levels of verb variation. Therefore, we created an entry model, which we call hub-node, to document such cases. It is aimed at representing the relation of the stable (or core) and variable idiom components. A depiction of the model with one example is displayed in Figure 5.

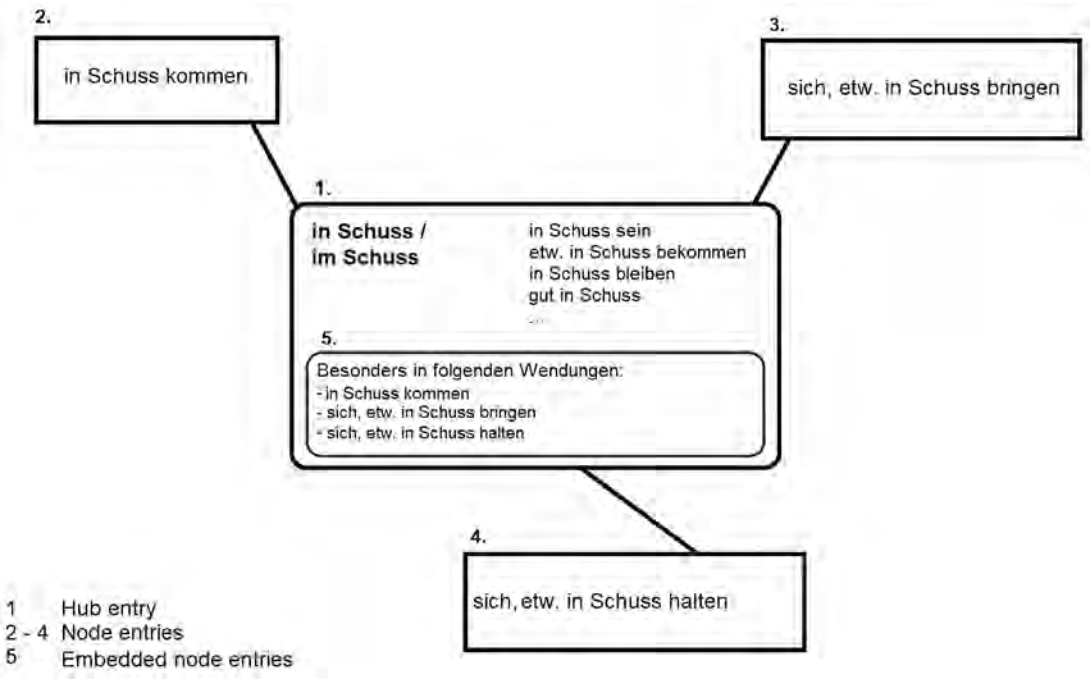


Fig. 5: Hub-node entry model

Hub entry In general, the hub entry such as *in Schuss/ im Schuss* (Fig. 5) meaning 'in a good state' serves as a central point in a (small) network of related entries. In particular, it serves as a container for less frequent variants of the idioms – variants which are typically not catered for in traditional dictionaries but are nonetheless present in corpora and might raise a need to consult the dictionary, e.g. *in Schuss sein* ('to be in a good state'), *etw. in Schuss bekommen* ('to get sth. in a good state'), *in Schuss bleiben* ('to stay in a good state') or *gut in Schuss* ('in a good state').

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

The main distinction of the hub entry from other MWE entry types is its lemma. In comparison to the usual MWE entries where the lemma is typically equal to an idiom (or to its canonical form known from other dictionaries) the hub entry is always titled with a phrase, which is the idiomatic nucleus or core metaphoric element of a number of the related idioms.

On the one hand, the reduction of the MWE lemma to its idiomatic nucleus allows the lexicographer to embed as many MWE variants as necessary in the dictionary entry; on the other hand, the idiomatic nucleus with its short and often underspecified form might look unfamiliar to the users. To eliminate this risk we decided to document the most common idioms, which the lemma of the hub entry is part of, as separate entries and describe them as node entries.

Node entry The primary function of the node entry is the lexicographic record of the salient and most common form of a verb idiom, which exhibits a high verb variation, under a distinct, easily recognisable lemma, in Figure 5: *sich, etw. in Schuss bringen* ('to bring oneself, sth. in a good state'), *etw. in Schuss halten* ('to keep sth. in a good state'), *in Schuss kommen* ('to come into a good state'). Node entries reduce the risk of the user not finding the commonly used forms of MWE.

Node entries follow the same micro-structure as the hub entry. They are fully-fledged entries of their own and can be consulted independently. Furthermore, they are projected into the hub entry with which they are associated.

Node entries are linked to the hub entry by a specialized link, which allows for the automatic projection of the content into the hub entry.

The node and the hub entry complement each other: the hub entry comprises the information not documented within the node entry and the node entry compensates for the short, often underspecified lemma of the hub entry.

With regard to its complexity, we consider the headword of a hub entry to be in between that of a single word headword and full idiomatic expression.

Synoptic entry As mentioned above, the content of the node entries is projected into the hub entry and forms a synopsis of them in it. In the DWDS-dictionary, it is positioned below the sense block. The synoptic section is headed by the phrase "Besonders in den folgenden Wendungen" ("mainly in the following expressions"). The function of this synopsis is a user-friendly representation of the common idiom forms in the hub entry and their easy accessibility of them through it.

In the following, we demonstrate the hub-node-model with the example of the hub entry *in Schuss/ im Schuss* ('in a good state') and one of the node entries *sich, etw. in Schuss bringen* ('to bring oneself, sth. into a good state') as well as the synoptic entry.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/ Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

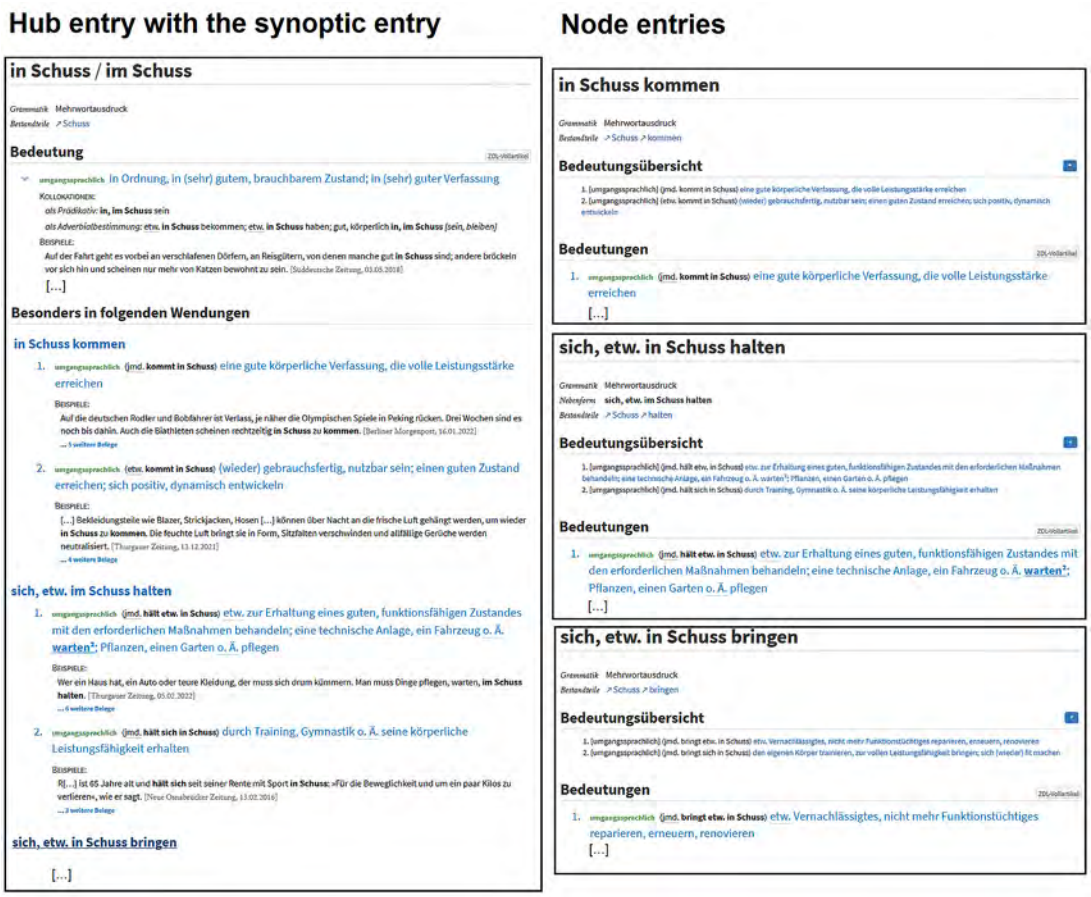


Fig. 6: Hub entry *in Schuss / im Schuss* (https://www.dwds.de/wb/in%20Schuss) and the respective node entries *in Schuss kommen* (https://www.dwds.de/wb/in%20Schuss%20kommen), *sich, etw. in Schuss halten* (https://www.dwds.de/wb/sich%2C%20etw.%20in%20Schuss%20halten), *sich, etw. in Schuss bringen* (https://www.dwds.de/wb/sich%2C%20etw.%20in%20Schuss%20bringen).

The model was successfully applied to other idiom groups, such as: hub entry *im / in den Keller* ('extremely low / at the very low point') with the node entries *im Keller sein* ('to be at rock bottom' about prices, state of mood), *in den Keller sinken* ('to fall on the lowest level'), *in den Keller gehen* ('to drop to the lowest level'); *Wind in den Segeln* ('wind in the sails' meaning '(new) motivation'), *Wind in die Segel blasen* ('to blow wind in the sails'), *Wind in den Segeln sein* ('to be wind in the sails').

Additionally, it should be mentioned, that the model is only applicable in those cases, where the nucleus is a multi-word expression itself.

More generally, the hub-node-entry model is applicable in cases where a) the headword of the hub entry can be described independently of the nodes, i. e. it is a (complex, multi-word) lexical unit with its own semantic core such as *bittere Pille* ('bitter pill' meaning 'sth. unpleasant') and b) there is a set of sense-related node entries which is neither too large nor too small.

It has to be decided on a case-by-case basis and only after a comprehensive consultation of the corpora (and other dictionaries) which way of treating a particular MWE is the most appropriate one. However, we see some potential of the hub-node-model beyond the example that has been presented in this paper.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

5. Conclusion and future work

In this article, we have introduced a model that helps to cope with the rich variation of use that comes along with many idiomatic expressions. Our model caters for both the less frequent but nonetheless relevant variants of an idiom (via hub entries) as well as with frequent variants (via node entries).

We try to follow a hybrid approach to lexicographically cover formal variation of various kinds of multi-word units:

1) most multi-word units are described in entries that are separate from the entries of their meaning-bearing components
2) the entries for the multi-word units are linked to the articles of their components and can be referenced from these component entries (see Fig. 2)
3) sets of multi-word entries with related meaning will be connected by lexical semantic cross-references in the same way as this is done in single word entries (e. g. synonymous senses, antonymous senses, associated senses, etc.)
4) multi-word expressions that share the same nucleus and exhibit a wide spectrum of variation can be described within the hub-node entry.

The next step for this approach will be to apply the hub-node entry model to other types of variation and modification. As future work, we plan to extend our work to other idiom patterns such as s. o. PREP ART N V where N is an open list of nouns, e. g. *jmdm. auf ART NN gehen* ('to get on someone's nerves (with something)) with an arbitrary range of lexical fillings for NN (*Nerven*, *Zeiger*, *Keks*, *Senkel* ...).

Among the central goals of the hub-node model is to give a much more comprehensive account of variation and modification of idioms in general language dictionaries in comparison to traditional dictionaries.

References

Dictionaries

Duden (2020): Duden 11. Redewendungen und sprichwörtliche Redensarten. Wörterbuch der deutschen Idiomatik. 5th edition. Berlin.

DUW = Deutsches Universalwörterbuch: Das umfassende Bedeutungswörterbuch der deutschen Gegenwartssprache. Hrsg. von der Dudenredaktion. 9th edition. Berlin 2019.

Schemann, H. (2011): Deutsche Idiomatik. Wörterbuch der deutschen Redensarten. Berlin/New York.

WDG = Wörterbuch der deutschen Gegenwartssprache. Hrsg. von Ruth Klappenbach und Wolfgang Steinitz. 6 volumes. Berlin, pp. 1964–1977.

WDW = WAHRIG Wörterbuch der deutschen Sprache. Hg. von Renate Wahrig-Burfeind. 27th edition. München 2018.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Other publications

Burger, H. (2015): Phraseologie. Eine Einführung am Beispiel des Deutschen. 5th edition. Berlin.

Elsen, H. (2017): Wortgruppenlexeme zwischen Wortbildung und Phraseologie. In: Yearbook of Phraseology 8 (1), pp. 145–160.

Fleischer, Wolfgang (1982): Phraseologie der deutschen Gegenwartssprache. Leipzig.

Hüning, M./Schlücker, B. (2015): „Multi-word expressions". In: Müller, P./Ohnheiser, I/Olsen, S./Rainer, F. (eds.): Word-formation. An international handbook of the languages of Europe. Berlin/New York, pp. 450–467.

Nunberg, G./Sag, I. A./Wasow T. (1994): Idioms. In: Language 70 (3), pp. 491–538.

Staffeldt, S. (2011): Die phraseologische Konstruktionsfamilie [X Präp *Hand* Verb]. In: Zeitschrift für germanistische Linguistik 39 (2), pp. 188–216.

Stumpf, S. (2019): Phraseografie und Korpusanalyse. In: Linguistik online 96 (3/19), pp. 115–131. https://bop.unibe.ch/linguistik-online/article/view/5523 (last access: 29-07-2021).

Contact information

Maria Ermakova
Zentrum für digitale Lexikographie der deutschen Sprache
Berlin-Brandenburgische Akademie der Wissenschaften
ermakova@bbaw.de

Alexander Geyken
Zentrum für digitale Lexikographie der deutschen Sprache
Berlin-Brandenburgische Akademie der Wissenschaften
geyken@bbaw.de

Lothar Lemnitzer
Zentrum für digitale Lexikographie der deutschen Sprache
Berlin-Brandenburgische Akademie der Wissenschaften
lemnitzer@bbaw.de

Bernhard Roll
Zentrum für digitale Lexikographie der deutschen Sprache
Berlin-Brandenburgische Akademie der Wissenschaften
roll@bbaw.de

Acknowledgements

The work in the project that has been funded by the German Ministry of Education and Research, (BMBF, FKZ: 01UG1902A). Special thanks goes to our colleagues who are currently working on dictionary entries for multi-word expressions and with whom we had lively discussions that finally led to the model presented here: Ralf Osterwinter, Lisa Palmes, Katrin Siebel, Stefanie Reckenthäler, Juliane Nau. We also thank Dmitrij Dobrovol'skij, Kathrin Steyer, Alexander Ziem and Carmen Mellado Blanco with whom we had the pleasure to discuss some of these issues in the Berlin-Workshop on MWE on Oct. 14th, 2021.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Semantics

XX EURALEX INTERNATIONAL CONGRESS
DICTIONARIES AND SOCIETY
12-16 July 2022, Mannheim, Germany

Robert Krovetz

AN INVESTIGATION OF SENSE ORDERING ACROSS DICTIONARIES WITH RESPECT TO LEXICAL SEMANTIC RELATIONSHIPS

Abstract This paper discusses an investigation of how senses are ordered across eight dictionaries. A dataset of 75 words was used for this purpose, and two senses were examined for each word. The words are divided into three groups of 25 words each according to the relationship between the senses: Homonymy, Metaphor, and Systematic Polysemy. The primary finding is that WordNet differs from the other dictionaries in terms of Metaphor. The order of the senses was more often incorrectly figurative/literal, and it had the highest percentage of figurative senses that were not found. We discuss leveraging another dictionary, COBUILD, to re-order the senses according to frequency.

Keywords Lexical semantics; word senses; corpus analysis

1. Introduction

The order of senses in a dictionary is an important problem in lexicography (Hiorth 1954; Kipfer 1983; Lew 2013). The literature usually discusses three orderings: 1) historical, 2) frequency, and 3) logical. The first order is used by the *Oxford English Dictionary* and the *Merriam-Webster* dictionaries.[1] The second order is used by dictionaries for learners of English as a second language, such as the *Longman Dictionary of Contemporary English*, *Oxford Advanced Learners Dictionary*, and *COBUILD*. There are differences about what is meant by the logical order, but literal/figurative, concrete/abstract, and general/specific are some of the distinctions that are mentioned. All dictionaries use a logical ordering to some extent.

There are several reasons why it is important to look at sense ordering. From the perspective of the user, we generally want the most frequent sense to be listed first. From the perspective of Computational Linguistics, we want to know sense order because it is important for word sense disambiguation. Many systems rely on the Most Frequent Sense (MFS) heuristic for classifying a word's sense in context (Agirre/Edmonds 2007). The skewed nature of word sense distributions means that the most frequent sense is not only more frequent, but often much more frequent than a secondary sense. From the perspective of lexicography, we want to get a better understanding of lexicographic judgment and how it relates to cognition. The frequency order in a corpus-based dictionary is not hard-and-fast. Learner's dictionaries will order senses in a way that is best for the user. There are also differences between historical order compared with logical order. For example, the figurative meaning for a word is sometimes older than the literal meaning.[2]

This paper discusses an investigation of how senses are ordered across eight dictionaries: *Longman Dictionary of Contemporary English* (LDOCE), *Collins English Dictionary* (COLLINS), *WordNet* (WN), *Cambridge International Dictionary of English* (CIDE), *Webster's*

[1] We note that the ordering is changing for the Merriam-Webster dictionaries so that the order would be what is most useful for the user.

[2] http://www.merriam-webster.com/words-at-play/6-words-whose-abstract-meanings-came-first/engine.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

New World Dictionary (WNW), *Oxford Advanced Learner's Dictionary* (OALD), *Collins COBUILD English Language Dictionary* (COBUILD) and *Merriam-Webster's 7th Collegiate Dictionary* (MW7). There were 75 words in the dataset, and two senses were examined for each word. There were 25 words for which the two senses were homonymous (e. g., *driver/car* and *driver/golf-club*, *draft/paper* and *draft/army*, *train/educate* and *train/locomotive*). Similarly, there were 25 words for which the relationship was a literal/metaphor difference (e. g., *shrimp/crustacean*, *shrimp/person*). There were instances within as well as across part-of-speech for these two subsets. For example, *train* was in the Homonymy dataset, and *parrot* was in the Metaphor dataset. Finally, there were 25 words that exhibited systematic polysemy between the senses. They were divided into different classes: animal/food, music/dance, language/people, tree/wood-of-tree, animal/hide, and natural-kind/color. All of these words were nouns. The dataset was created by the author for a variety of purposes, and it is a subset of a larger sense inventory. For words that differed in part-of-speech, we looked at the order of homographs.

The study had the following aims:

1) Compare three different learner's dictionaries to see where they agree, and where they differ, in the ordering of the senses in the three datasets. We used the *Longman Dictionary of Contemporary English* (LDOCE), *Oxford Advanced Learner's Dictionary* (OALD), and *Collins COBUILD English Language Dictionary* (COBUILD).
2) Compare WordNet[3] against the other dictionaries in the same regard. Because of the importance of this dictionary as the basis for the Most Frequent Sense (MFS) heuristic for word-sense disambiguation, we wanted to see how easy it is to leverage the other dictionaries to re-order the senses in WordNet when senses are in the wrong order.
3) Compare a Merriam-Webster dictionary (MW7) against the other dictionaries to see how historical order differs.
4) Examine cases where there is disagreement and assess which senses are most frequent. This was done by using ngrams that occur in different corpora: Project Gutenberg (gutenberg.org), the Wikipedia (en.wikipedia.org), and the Internet, as represented by the Google n-grams dataset (Brants/Franz 2006). A sample of the ngrams is given in Table 1. These ngrams contained the target word (or an inflected form), and they were manually selected. Multiple ngrams were reviewed for each word in order to make the assessment. These ngrams were also used in assessing the order of senses in WordNet. The use of ngrams for this purpose is based on the *One Sense per Collocation* hypothesis (Yarowsky 1990).

The next section will discuss the results of the comparison, and a small experiment to re-order the senses in WordNet.

Word	**Semantic Relation**	**Ngram pair**
boom	homonymy	booming voice/booming economy
discount	homonymy	discounted rate/discount the importance
draft	homonymy	draft a bill/I was drafted
entitle	homonymy	entitled to vote/entitle an act
frisk	homonymy	frisk and play/stop and frisk
gag	homonymy	running gags/mouth gag

[3] We used version 3.0

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Word	Semantic Relation	Ngram pair
john	homonymy	prostitutes and johns/use the john
mortar	homonymy	mortar shells/mortar and pestle
stigma	homonymy	social stigma/stigma and pollen
bask	metaphor	bask in the sun/bask in the glory
defuse	metaphor	defuse the bomb/defuse tension
postmortem	metaphor	postmortem exam/postmortem analysis
conceive	metaphor	children conceived/ill conceived
purgatory	metaphor	souls in purgatory/kind of purgatory
shrimp	metaphor	brine shrimp/bully and the shrimp
underline	metaphor	words underlined/underline the importance
uproot	metaphor	uproot the trees/uproot themselves

Table 1: A sample of the ngrams used to assess sense ordering

2. Results

Table 2 gives the results of the comparison, and which sense was found to be most frequent according to the ngram analysis.

The literature mentions difficulty in making comparisons between the senses in different dictionaries (Atkins/Levin 1991), (Kilgarriff 1997). In contrast, we found it was fairly straightforward to identify the senses in the dataset and how they were ordered in the different dictionaries. This is because we are looking for specific senses rather than a general mapping.

The main problematic dictionary was CIDE. It did not enumerate the senses like the other dictionaries, but rather provided bullets, which were also used for example sentences. This made it difficult to determine when a sense was being distinguished. The main sense mapping problem was with the word *sandwich*. We were looking for a figurative sense involving time (e. g., *I can sandwich you in between 2 and 3 PM*). Most of the dictionaries defined the figurative sense only in terms of space.

Part of speech was sometimes a problem. Dictionaries differed in whether a word sense was attested as a noun or as an adjective (e. g., *turquoise*). We allowed such differences in matching the sense we were looking for.

Morphological variation was a factor, and sometimes the sense was found only under a variant form (e. g., *inflect* vs. *inflection*). In addition, sometimes the sense associated with a word form was not found, such as *plastered*, which can either mean "apply plaster" or "drunk".

The results for the different lexical semantic classes are given next. This is followed by a discussion of WordNet and a small experiment at re-ordering the senses.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Word	Label	Dictionaries	Corpus Results
boom	sound:grow-rapidly	COBUILD, CIDE	CORPUS-SPECIFIC
discount	bargain:opinion	WN,OALD, COLLINS, LDOCE	bargain
draft	army:paper	LDOCE,WN,COBUILD,OALD	paper
entitle	book:permitted	COLLINS,COBUILD,WN	permitted
frisk	playful:search	COBUILD,OALD	CORPUS-SPECIFIC
gag	mouth:joke	WN	CORPUS-SPECIFIC
inflect	word:voice	COBUILD,WNW	word
john	prostitute-client:toilet	(LDOCE,CIDE)/(WN,WNW)	toilet
mortar	pestle:gun	all except WNW	gun
stigma	shame:plant	WN	shame
bask	sun:approval	WN	sun
conceive	baby:imagine	COLLINS,WN,COBUILD	imagine
defuse	bomb:situation	(LDOCE,OALD)/COBUILD	situation
postmortem	death:final-analysis	WN	death
purgatory	hell:bad-place	WN	hell
shrimp	crustacean:person	WN	crustacean
underline	writing:emphasis	COBUILD, WN	CORPUS-SPECIFIC
uproot	plants:from-home	WN,COBUILD	CORPUS-SPECIFIC

Table 2: Words where there was disagreement about the order of the senses compared with historical order. The table shows the sense labels in historical order (where available), the dictionaries where the order was different, and the sense that was most frequent according to corpus analysis. The first part of the table illustrates words that are in the Homonymy dataset, and the second part shows words in the Metaphor dataset

2.1 Homonymy and Metaphor Datasets

There were 5 words each in the Homonymy and Metaphor datasets that differ in part-of-speech between the senses. These were usually represented as different homographs in the dictionaries and almost all were ordered the same way. The only exception was *novel*, which was ordered adjective first except for COBUILD and Collins.

Of the 20 remaining words in the Homonymy dataset, there were differences in ordering for 10 words. For the Metaphor dataset, 8 of the words differed in sense order. The three learner's dictionaries (OALD, LDOCE, COBUILD) differed in their ordering for 5 out of 20 words in the Homonymy dataset, and for 4 out of 20 in the Metaphor dataset. These words are given in Table 2.

There was one word, *john*, which was only listed in one sense (toilet) in MW7, so the disagreement is given for the other dictionaries. Except for *john* and *defuse*, the order of senses in the Label column is the historical order.[4]

[4] The order for *john* in the Labels column (prostitute-client:toilet) corresponds to the difference between the four dictionaries, which were the only ones in which both senses were listed. The word *defuse* was not defined in MW7.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

2.2 Systematic Polysemy Dataset

For the Systematic Polysemy dataset there was inconsistency within as well as between dictionaries. Words in each of the subsets (animal/food, music/dance, language/people, tree/wood-of-tree, animal/hide, natural-kind/color) were sometimes in that order, and sometimes not, depending on the dictionary and the word.

It is not surprising that the order differs for this set. Nor do we feel that the order is necessarily important for the users of the dictionary. It is more important in Computational Linguistics, where the distinctions are needed for natural language understanding. Our main interest in this set is from a cognitive perspective. (Panman 1982) observed that when word senses are homonymous, people will agree that they are different. But when the senses are related, people will disagree about whether the senses are distinct. We wanted to look at this question from the perspective of lexicographic judgment. How often are senses in this set distinguished compared with the other datasets? That is, to what extent does (Panman 1982)'s comment apply to lexicographers, especially since they are considered to be splitters rather than lumpers (Bejoint 1988).

The Homonymy class was indeed individuated more often than the Systematic Polysemy class. However, there were significant differences between dictionaries and between words within a class:

1) The Cambridge Dictionary did not include the Language/People class, and distinguished the senses of only 5 out of 25 words that were systematically polysemous. Webster's New World Dictionary distinguished all 25. The rest of the dictionaries distinguished 18 or more.
2) The Natural-Kind/Color class (*gold*, *silver*, *jade*, *rust*, *turquoise*) was distinguished most often, and the Music/Dance class (*waltz*, *tango*, *foxtrot*, *rumba*, *polka*) was distinguished least often amongst the Systematic Polysemy classes. The Tree/Wood-of-tree group was distinguished most often for *oak* and *pine*, and less often for *maple* and *chestnut*. This is in accord with (Hanks 1979) and many others, who have noted that dictionaries contain senses that are stereotypical of usage.
3) The Metaphor class was in between homonymy and systematic polysemy. Dictionaries differed in whether the distinction was made in a separate homograph, in a separate sense within a homograph, or as a subsense. They also differed on whether the sense was labeled as figurative.

2.3 Frequency, learner's dictionaries, and historical order

We used ngrams from three large corpora (Project Gutenberg, the Wikipedia, and the Google n-grams dataset) to make an assessment of the ordering based on corpus frequency, and the sense which is most frequent is indicated in Table 2.

The corpus results generally support the ordering in the learner's dictionaries for the Homonymy dataset. The corpus results for *discount* differed from most of the learner's dictionaries, and the ordering in MW7 was in agreement with the corpus data. The results for *inflect* depended on the word form. The root was typically associated with inflecting a word (the older sense), but the derived form *inflection* was associated more often with vocal inflection, and *inflectional* was associated with inflecting a word. For all words in the Metaphor dataset, MW7 ordered the metaphorical extensions after the literal meaning, which is generally what we would expect, and so there is less disagreement for the Metaphor dataset.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

That is, the historical order and the most frequent order are generally the same. An exception is the words *conceive* and *defuse*, which were more frequent in the figurative than the literal sense in all three corpora.

2.4 WordNet and Metaphor

The results for the Metaphor dataset showed that the literal sense was usually most frequent, but not always, and the most frequent sense is sometimes corpus-specific (e.g, *underline* is used most frequently in the literal sense in the Project Gutenberg corpus, and the figurative sense is most common in the Wikipedia and the Internet-based corpus).

WordNet differed from all of the other dictionaries in terms of Metaphor. It had the most words (8 out of 20) that were only defined in the literal sense. WordNet also ordered metaphorical senses before literal senses more often than most of the other dictionaries (6 words compared with 4 words).

WordNet is the most widely used dictionary in Computational Linguistics, and the ordering is based on the frequency of the senses in SemCor, a subset of the Brown corpus that has been manually tagged with senses from WordNet (Landes/Leacock/Tengi 1998). This corpus is small (only about 200,000 word tokens), and many word senses appear infrequently. For example, the word *tiger* appears only three times. Two of the tokens were used in a figurative sense, and this was used to order the figurative sense before the literal one.

We conducted a small experiment to leverage the ordering in COBUILD to re-order the corresponding senses in WordNet with regard to a literal/metaphor distinction. This was based on a manual mapping between the senses in our own sense inventory, and the corresponding senses (where found) in WordNet We looked at an additional 20 words that have a literal/figurative distinction: *agitator*, *avalanche*, *bankrupt*, *barrage*, *beak*, *beanpole*, *bloodsucker*, *blight*, *lamb*, *leech*, *pedestal*, *shark* (nouns), and *applaud*, *backfire*, *backtrack*, *bait*, *devour*, *nosedive*, *unmask*, *unseat* (verbs). Of these words, five were defined only in a literal or figurative sense in WordNet, and one word was not defined. Most of the remaining words were in a literal/figurative order in both WordNet and in COBUILD. We examined those senses where the figurative sense was listed before the literal sense. As Table 2 shows, there are some cases such as *conceive* and *defuse* where a figurative sense is more frequent than a literal sense, but for most of the words in our dataset the literal sense is more frequent. We then examined COBUILD to see if the order supported the ordering in WordNet, or if the ordering differed. The aim is to leverage the larger corpus frequency that COBUILD is based on. The ordering differed for *blight*, *bait*, and *devour*. *We* identified ngrams in the Project Gutenberg, Wikipedia, and Google datasets for these three words. We found that *bait* and *blight* are generally used literally more often than figuratively,[5] but *devour* was often used in both a figurative and literal sense and the order depended on the corpus.

3. Discussion

The order of senses is not an easy decision. There can be a conflict between the most frequent sense and the most salient sense, or between the most frequent sense and a sense

[5] The primary exceptions are *bait and switch*, *urban blight*, and *suburban blight*.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

order that would follow a consistent pattern such as putting the literal sense before the figurative one.

From the perspective of Computational Linguistics though, the decision is easier – we want to identify the sense that is most frequent. The Most Frequent Sense heuristic is widely used in research on word sense disambiguation (Agirre/Edmonds 2007). It is used as a back-off method when we do not have enough information to make a more informed choice. We found it was relatively straightforward to identify corresponding senses between WordNet and other dictionaries with regard to a literal/figurative distinction, and we were able to use this information to propose a re-ordering of the WordNet senses. We were also able to use corpus ngrams from Gutenberg, the Wikipedia, and the Internet to support that re-ordering.

The more difficult problem is the missing senses in WordNet. It stood out among the dictionaries in terms of identifying figurative senses least often. In the sample of 20 words that we used to assess sense order and metaphor, five of the words were missing a literal or a figurative sense in WordNet, and one word was not defined. WordNet has been criticized for being „too fine-grained" (making too many distinctions), but this is a case where additional distinctions are needed.

The results on word-sense individuation show that there is a great deal of consistency for words that are in the Homonymy class. Over the set of of 25 words, we found that almost all dictionaries distinguished the senses. The dictionaries differed in the order of the senses, but usually not in the fact that they were distinguished. The results on the Systematic Polysemy dataset show that the senses in this group are distinguished least often, and this is what we would expect. (Panman 1982) found that when two senses are homonyms, people agree that the senses are different, and when the senses are related people disagree about whether they are distinct meanings. However, we found that even amongst the sets of words that are systematically related, there is an ordering of different to similar. Dictionaries distinguished senses most often for words with a substance/color relationship (as with *gold*, *silver*, and *amber*), and least often for words with a music/dance relationship (as with *waltz*, *foxtrot*, and *tango*).

4. Conclusion

This paper looked at sense-ordering across a number of different dictionaries. WordNet differed from all the dictionaries with respect to Metaphor. A small experiment showed that the COBUILD dictionary and an ngram analysis can be leveraged to re-order those senses that were out-of-order with regard to frequency. The dataset of 75 words and the spreadsheet about their ordering in the different dictionaries is available from: http://lexicalresearch.com/resources/euralex-2022-dataset.tar

References

Agirre, E./Edmonds, P. (eds.) (2007): Word sense disambiguation: algorithms and applications, Heidelberg.

Atkins, B./Levin, B. (1998): Admitting impediments. In: Zernik U. (ed.): Lexical acquisition: exploiting on-line resources to build a lexicon. Hillsdale, NJ.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Bejoint, H. (1988): Monosemy and the dictionary. In: BudaLex'88 Proceedings: Papers from the 3rd International Euralex Congress.

Brants, T./Franz, A. (2006): Web 1T 5-gram corpus version 1.1. Technical report, google research. The resource is available from the linguistic data consortium. https://catalog.ldc.upenn.edu/LDC2006T13.

Cambridge international dictionary of English (1995). Cambridge, MA.

Collins COBUILD English language dictionary (1987). London/New York.

Collins English dictionary (1979). First edition. London/New York.

Fellbaum, C. (ed.) (1998): WordNet: an electronic lexical database. Cambridge, MA.

Hanks, P. (1979): To what extent does a dictionary definition define? In: ITL International Journal of Applied Linguistics 45, pp. 32–38.

Hiorth, F. (1954): Arrangement of meanings. In: Lexicography, Lingua 4, pp. 413–424.

Kilgarriff, A. (1997): I don't believe. In: Word Senses, Computers and the Humanities 31 (2), pp. 91–113.

Kipfer, B. (1983): Methods of ordering senses within entries. In: Proceedings of Euralex, pp. 49–54.

Landes, S./Leacock, C./Tengi, R. (1998): Building semantic concordances. In: Fellbaum, C. (ed.): WordNet: an electronic lexical database. Cambridge, MA, pp. 199–216.

Lew, R. (2013): Identifying, ordering and defining senses. In: The Bloombury companion to lexicography, pp. 284–302.

Longman dictionary of contemporary English (1978). London.

Oxford advanced learner's dictionary (1989). Fourth edition. Oxford.

Panman, O. (1982): Homonymy and polysemy. In: Lingua, pp. 105–136.

Webster's seventh new collegiate dictionary (1965). Springfield, MA.

Webster's new world dictionary of the American language (1970). Second college edition. New York.

Yarowsky, D. (1993): One sense per collocation. In: Proceedings of the Workshop on Human Language Technology, pp. 266–271.

Contact information

Robert Krovetz
Lexical Research
rkrovetz@lexicalresearch.com

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Index of Authors

INDEX OF AUTHORS

A

Abel, Andrea 449
Aldea, Maria 650
Alves, Ieda Maria 804
Arapopoulou, Maria 725

B

Bajčetić, Lenka 387
Ballestracci, Sabrina 460
Bartels, Hauke 540
Bichlmeier, Harald 660
Bielinska, Monika 690
Brač, Ivana 334

C

Chambat, Anaïs 735
Choi, Jun 814
Costa, Rute 181

D

Declerck, Thierry 296
De Martin Pinter, Patrizio 99
de Schryver, Gilles-Maurice 196, 833
Diewald, Nils 208
Dimou, Athanasia-Lida 357
DiMuccio-Failla, Paolo 99
Domínguez Vázquez, Maria Josè 690
Dorn, Nico 368
Doupas, Athanasios 625
Ducassé, Mireille 381
Dzhuranyuk, Vova 401

E

Efthimiou, Eleni 357
Elizbarashvili, Archil 381
Engelberg, Stefan 87
Ermakova, Maria 851

F

Flinz, Carolina 460
Flouda, Christina 625
Fotinea, Stavroula-Evita 357
Füreder, Birgit 301

G

Gantar, Polona 240, 549
Garoufos, Apostolos 563
Gavriilidou, Zoe 471, 563, 614
Geyken, Alexander 851
Giacomini, Laura 99
Gînsac, Ana-Maria 222
Giouli, Voula 625
Gloning, Thomas 23
González Ribao, Vanessa 253, 569
Goulas, Theodoros 357
Gouws, Rufus H. 36, 690
Grønvik, Oddrun 321

H

Harm, Volker 701
Heid, Ulrich 480
Hollós, Zita 436

I

Ilić, Velibor 387

J

Johannsson, Ellert Thor 777
Jung, Hae-Yun 814

K

Kalafikis,Georgios 725
Kallas, Jelena 509
Karamitsou, Dimitra 725
Kernerman, Ilan 401
Klaes, Christiane 310
Klosa-Kückelhaus, Annette 113
Koeva, Svetla 509
Konstandinidou, Evi 471
Kosem, Iztok 230, 240, 509
Kouassi, Konan 172
Krek, Simon 240, 549
Krovetz, Robert 862
Kruse, Theresa 480
Kupietz, Marc 208
Kupriianov, Yevhen 584

L

Labropoulou, Penny 310
Langemets, Margit 509
Langer, Gabriele 635
Lemnitzer, Lothar 851
Lindemann, David 310
Lonke, Dorielle 401
Lorentzen, Henrik 825
Lüngen, Harald 208

M

Makino, Takahiro 409
Maroneze, Bruno 804
McLelland, Nicola 53
Meliss, Meike 253
Meyer, Peter 578
Michaelis, Frank 346
Minde, Trond 321
Miyata, Rei 409

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Moruz, Mihai-Alex 222, 745
Moshövel, Andrea 711
Müller, Anke 635
Müller-Spitzer, Carolin 129

N

Nabirye, Minah 833
Nazar, Rogelio 262
Nied Curcio, Martina 71, 690

O

Ostapova, Iryna 584
Ostroški Anić, Ana 334
Otte, Felicitas 635

P

Panocová, Renáta 792
Pavlova, Anna 594
Petrović, Snežana 387
Pinnavaia, Laura 142
Plate, Ralf 605
Pozzi, María 678
Proost, Kristel 346

Q

R

Renau, Irene 262
Roll, Bernhard 851
Rüdiger, Jan Oliver 129, 346

S

Salgado, Ana 181, 423
Salveridou, Kyriaki 614
Sarischoulis, Efstratios 725
Sato, Satoshi 409
Schierholz, Stefan J. 690
Shyrokov, Volodymyr 584
Sidiropoulos, Nikos 625
Simões, Alberto 423
Smith, Chris A. 273
Smith Ore, Christian-Emil 321
Španović, Ana 387
Stainhaouer, Gregory 625
Stincone, Clarissa 755
Storjohann, Petra 155
Sungwon, Seo 409
Sviķe, Silga 494

T

Tasovac, Toma 181
ten Hacken, Pius 792
Tiberius, Carole 509
Trap-Jensen, Lars 825
Tselikas, Sotiris 725

U

Ungureanu, Mădălina 222, 745

V

Vacalopoulou, Anna 357, 625
Vasilaki, Kiki 357

W

Wähl, Sabrina 635
Wigestrand Hoftun, Agnes 522

X

Y

Yablochkov, Mykyta 584

Z

Žarković, Marija 765
Zeschel, Arne 346

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.