

<http://lexikos.journals.ac.za>; <https://doi.org/10.5788/35-2-2108> (Volle uitgawe / Full volume)

Lexikos 35(2)

<http://lexikos.journals.ac.za>; <https://doi.org/10.5788/35-2-2108> (Volle uitgawe / Full volume)

Lexikos 35(2)

Hoofredakteur / Editor-in-Chief

André H. du Plessis

Mederedakteurs / Associate Editors

Dion Nkomo

Elsabé Taljard

Michele van der Merwe

Sonja Bosch (Duits/German)

Steve Ndinga-Koumba-Binza (Frans/French)

Resensieredakteur / Review Editor

Dané Claassen



African Association for Lexicography

AFRILEX-REEKS 35A:2025

AFRILEX SERIES 35A:2025



BURO VAN DIE WAT

STELLENBOSCH

<http://lexikos.journals.ac.za>; <https://doi.org/10.5788/35-2-2108> (Volle uitgawe / Full volume)

Uitgewer Publisher

BURO VAN DIE WAT
Posbus 245
7599 STELLENBOSCH

Kopiereg © 2025 deur die uitgewer
Alle regte streng voorbehou
Eerste uitgawe 2025

Tipografie en uitleg deur Maryna Adshade, Tanja Harteveld en Hermien van der Westhuizen
Bandontwerp deur Piet Grobler
Geset in 10 op 12 pt Palatino

ISBN 978-0-6398803-0-3
ISSN 2224-0039

Hierdie werk is gelisensieer ingevolge 'n Creative Commons License CC BY 4.0-lisensie.

Licensed under Creative Commons License CC BY 4.0.

Menings wat in artikels en resensies uitgespreek word, is nie noodwendig dié van AFRILEX of die Buro van die WAT nie.

Opinions expressed in the articles and reviews are not necessarily those of AFRILEX or of the Bureau of the WAT.

Lexikos is elektronies beskikbaar by <http://lexikos.journals.ac.za/>

Lexikos is available online at <http://lexikos.journals.ac.za/>

Lexikos is elektronies beskikbaar by Sabinet, AJOL, Ebsco en Proquest

Lexikos is available online from Sabinet, AJOL, Ebsco and Proquest

Indekse Indexes

Arts and Humanities Citation Index® (Clarivate); Asian Digital Library; Current Contents®/ Arts & Humanities (Clarivate); Current Contents®/Social and Behavioral Sciences (Clarivate); ERIH Plus; EuroPub Index; Index Copernicus Journals Master List; Journal Citation Reports/Social Sciences Edition (Clarivate); Linguistic Bibliography Online (Brill); Linguistics Abstracts Online; Linguistics and Language Behavior Abstracts; MLA International Bibliography; R.R.K. Hartmann's Bibliography of Lexicography; SciELO SA; Scopus (Elsevier); Social Sciences Citation Index® (Clarivate); Social Scisearch® (Clarivate)

Adviesraad / Advisory Board

Prof. H.L. Beyer (Namibië/Namibia)
Prof. G.-M. de Schryver (België/Belgium)
Prof. M.J. Domínguez Vázquez (Spanje/Spain)
Prof. P.A. Fuertes-Olivera (Spanje/Spain)
Prof. R.H. Gouws (RSA)
Prof. R. Lew (Pole/Poland)
Prof. T. Margalitadze (Georgië/Georgia)
Prof. P.A. Mavoungou (Gaboen/Gabon)
Prof. D.J. Prinsloo (RSA)
Prof. S. Tarp (Denemarke/Denmark)
Prof. H. Xu (China)

Redaksiekomitee / Editorial Committee

Prof. N. Bosman (RSA)
Dr. E.M. Ella (Gaboen/Gabon)
Prof. Y. Gao (China)
Prof. L. Khumalo (RSA)
Dr. P.A. Louw (RSA)
Dr. L. Mbindi Aninga (Gaboen/Gabon)
Dr. L. Morris (RSA)
Prof. E. Ndlovu (Zimbabwe)
Dr. G. Odendaal (RSA)
Dr. R.N. Ossai (Nigerië/Nigeria)
Dr. A. Ostroški Anić (Kroasië/Croatia)
Prof. T.J. Otlogetswe (Botswana)
Prof. A.N. Otto (RSA)
Dr. S. Wolfer (Duitsland/Germany)

Inhoud / Contents

Voorwoord	ix
Foreword	xii
<i>André H. du Plessis</i>	
'n Woord van AFRILEX	xv
A Few Words from AFRILEX	xvi
<i>Langa Khumalo</i>	
Redaksionele doelstellings	xvii
Editorial Objectives	xviii

Artikels / Articles

Die gebruik van kleur in aanlyn woordeboeke, met oorweging van drie Afrikaanse woordeboeke	1
<i>Maryna Adshade</i>	
Thematic Classification in Polish Electronic Dictionaries — From Tradition to New Tools and Opportunities	302
drie Afrikaanse woordeboeke	
<i>Barbara Batko-Tokarz</i>	
Generative Artificial Intelligence (GenAI) as Information Tool for Lexicographic Information Needs	276
<i>Theo J.D. Bothma and Rufus H. Gouws</i>	
Gender Bias in Computer-generated Thesauri: The Case of the Serbian Section of <i>Kontekst.io</i> , a Thesaurus of Synonyms and Semantically Related Terms	26
<i>Dragana Čarapić and Milica Vuković-Stamatović</i>	
Out-of-the-Box GPTs for Lexicography	362
<i>Gilles-Maurice de Schryver</i>	
Customised GPTs for Lexicography	397
<i>Gilles-Maurice de Schryver</i>	

'n Omvattender aanbod van semantiese data in aanlyn woordeboeke <i>Rufus H. Gouws</i>	46
Lexicographic Description of a Polysemous Word in a Learner's Dictionary Based on Its Lexical Prototype <i>Anton Karasev</i>	75
Exemplification of Sensitive Words for People with Disabilities in Monolingual English Learner's Dictionaries <i>Meimei Li</i>	94
Lexicography in Action: The Traversal from Coinage and Iconicity to Iconisation <i>Dorra Moalla Masmoudi</i>	117
Reconceptualising Contemporary Lexicography: Emotional Intelligence and the Sustainable Lexicographic Modelling of the Modal Verb <i>moći</i> <i>Milena Mrdak-Mićović and Nataša Jovović</i>	143
Aktuelle Aussprachewörterbücher des Deutschen und des Englischen im Vergleich <i>Krzysztof Nycz und Zygmunt Tęcza</i>	232
Spelling Variations in <i>A Dictionary of Setswana Personal Names</i> <i>Thapelo J. Otlogetswe and Goabilwe Ramaeba</i>	324
Making Dictionary Content Accessible for People with Visual Impairments <i>Geraint Paul Rees</i>	165
One Name, Two Genres: The Curious Case of Polish Encyclopaedias of Law <i>Agnieszka Rzepkowska</i>	185

Projekte / Projects

Where Tools Are Few: Constructing Limited-Word Bilingual Learner's Dictionaries in a Low-Resourced Community in Malawi <i>Ian David Dicks</i>	256
--	-----

Lexicographic Post-processing and Refinement Issues in Wordnets: The Ukrainian Wordnet Experience <i>Maksym Vakulenko</i>	345
---	-----

Leksikonotas / Lexiconotes

A Systematic Literature Review of Over Three Decades of Dictionary Use Research in <i>Lexikos</i> <i>Balázs Fajt, Mátyás Bánhegyi and Katalin P. Márkus</i>	423
---	-----

Noms de Marques dans la 9 ^e Edition du <i>Dictionnaire</i> <i>de l'Académie Française</i> <i>Jan Holeš et Zuzana Honová</i>	215
--	-----

Leksikohuldeblyk / Lexicotribute

Frederik Johannes Lombard 1953 – 2025 <i>Alet Cloete, Phillip Louw en/and Gerhard van Wyk</i>	296
--	-----

Publikasieaankondigings / Publication Announcements

Voorskrifte aan Skrywers	444
Instructions to Authors	445

Voorwoord

In 1989 het 'n formele leksikografiese gesprek op Stellenbosch tussen die redaksie van die *Woordeboek van die Afrikaanse Taal* (WAT) en 'n aantal gerekende taalkundiges plaasgevind. Hierdie sleutelfigure het besef dat daar 'n groot behoefte aan die uitruil van leksikografiese kennis en ondervinding in Afrika bestaan. Dit het uiteindelik gelei tot die totstandbrenging van *Lexikos*, 'n uitstekende medium en voertuig vir leksikografiese gesprekvoering. *Lexikos* 1, 'n stuk leksikografiese pionierswerk, verskyn in 1991 onder die stuur van wyle mnr. Pieter Harteveld. Vir die eerste keer het Afrika 'n vaktydskrif wat hoofsaaklik op die leksikografie gerig is. Hierdie tydskrif het 35 jaar later kennelik een van die vernaamste internasionale vaktydskrifte vir en oor die leksikografie geword. Die menigte toonaangewende *Lexikos*-bydraes oor die laaste drie en 'n half dekades het 'n blywende en beduidende impak op die leksikografie as vakrigting gehad. Die redaksie, uitgewer (die WAT) en AFRILEX is geweldig trots op die produk wat oor die afgelope 35 jaar uitgegee is. Mag *Lexikos* vir die volgende 35 jaar die leksikografie ter wille wees en mag daar nog vele volumes suksesvol verskyn!

Soos met die publikasie van *Lexikos* 1, het die 35ste volume ook vir uitdagings en moeilike besluite gesorg. Sedert 2022 het die aantal bydraes wat ingelewer is, stelselmatig vermeerder, maar in 2025 het die spreekwoordelike sluise behoorlik oopgetrek. Die redaksie het watergetrap terwyl daar 'n plan geformuleer moes word om al die bydraes te verwerk en te publiseer. Daar is uiteindelik, met verlot van die uitgewer, besluit om vanjaar twee uitgawes te publiseer. Hierdie geweldige toename in bydraes getuig van die reikwydte en aansien van *Lexikos*. Dit is verblydend dat so baie outeurs *Lexikos* as geskikte tuiste vir hul navorsingsuitsette ag. Dit plaas egter ongewone druk op die redaksie, redaksionele steundienste en uitgewer. Die toename in artikels gepaard met *Lexikos* se huidige redaksionele en finansiële model het aanleiding gegee tot sinvolle en nodige gesprekke tussen AFRILEX, die redaksie en die uitgewer oor die volhoubare bestuur van die tydskrif. Die besluit is solank geneem dat *Lexikos* 35 die laaste volume sal wees wat bladgeld slegs van Suid-Afrikaanse outeurs eis. Vanaf 2026 sal die eis van publikasiefoeie ("article processing charges") van outeurs buite Suid-Afrika geïmplementeer word. *Lexikos* se missie is wel om steeds te verseker dat kennis gratis met die leksikografiese gemeenskap gedeel word. Derhalwe bly *Lexikos* 'n ooptoegangtydskrif vir alle lesers.

Vanjaar se eerste uitgawe bevat 'n spesiale afdeling wat toegewy is aan die *6th GLOBALEX Workshop on Lexicography and Neology*. Hierdie werkswinkel is met groot sukses by die *28th International Conference of the African Association for Lexicography* aangebied. Sewe uit die elf referate is in artikels omskep en bied interessante en unieke perspektiewe op die skep, gebruik en leksikografiese ontginning van neologismes die wêreld oor. Ons is dankbaar vir die samewerking

met GLOBALEX en waardeer die redaksionele en koördineringswerk wat die gasredakteurs van hierdie afdeling, prof. Annette Klosa-Kückelhaus en mnr. Ilan Kernerman, behartig het.

Die ander bydraes in vanjaar se uitgawes is, soos verwag, verteenwoordigend van die verskeidenheid navorsingmoontlikhede in die leksikografie en terminologie. Daar is insiggewende en waardevolle bydraes oor onder andere terminologiebestuur, kunsmatige intelligensie, aanlyn woordeboeke, die metaleksikografie en inklusiewe leksikografie. Naas ons outeurs uit verskillende dele van Afrika, soos Botswana, Gaboen en Tunisië, is daar stemme uit, onder andere, Australië, China, Litaue, Montenegro, Pole en Spanje. Baie dankie aan elke outeur wat deur middel van hul publikasies in *Lexikos* aan die leksikografiese diskoers deelneem. Die redaksie is ewig dankbaar vir al die outeurs se samewerking en positiewe gesindhede.

Die publisering van *Lexikos* is altyd 'n spanpoging en die span wat 'n aandeel in vanjaar se uitgawes gehad het, is uiters bekwaam. 'n Spesiale woord van dank aan proff. Elsabé Taljard, Dion Nkomo en Michele van der Merwe vir die uitstekende werk wat hulle as mederedakteurs gedoen het, asook vir die waardevolle raad en bystand wat hulle gebied het. Baie dankie ook aan dr. Steve Ndinga-Koumba-Binza en prof. Sonja Bosch vir hulle volgehoue en uiters knap werk as Franse en Duitse redakteurs. Baie dankie aan dr. Dané Claassen vir die uitsonderlike werk met die resensies. Dit is 'n voorreg om op so 'n behendige redaksie te kan staatmaak. Aan mee. Tanja Harteveld en Hermien van der Westhuizen van die WAT, wil ek namens die redaksie opreg dankie sê vir hulle toewyding aan die stiptelike en professionele set en afronding van vanjaar se uitgawes, die administratiewe bystand, asook vir hulle bekwaamheid met al die ad hoc-take wat deur die loop van die jaar opgeduik het. So ook bedank ek namens die redaksie ons nuutste spanlid, me. Maryna Adshade. As redaksionele assistent het me. Adshade haar ontsettend goed van haar taak gekwyt en ons is innig dankbaar vir haar bereidwilligheid, flinkheid, toewyding en ondersteuning. 'n Hartlike bedanking ook aan dr. Phillip Louw, Uitvoerende Direkteur van die WAT, vir sy toegemoetkomendheid en toewyding aan *Lexikos*.

Lexikos is groot dank verskuldig aan die kranige portuurbeoordelaars wat verseker het dat vanjaar se bydraes die hoë wetenskaplike standaard van hierdie tydskrif handhaaf.

Ons het vanjaar met groot leedwese verneem van die heengaan van dr. Frikkie Lombard, voormalige eindredakteur van die WAT. Dr. Lombard het diep spore in die Suid-Afrikaanse, veral die Afrikaanse leksikografie gelaat. Dr. Lombard het 42 jaar die WAT met onderskeiding gedien, en 25 jaar in die Taalkommissie. Hy het in daardie tyd 'n blywende merk op die Afrikaanse taallandskap gemaak, veral deur die 10 WAT-dele waarby hy betrokke was en die betekenisvolle bydrae wat hy gemaak het om die Afrikaanse spelreëls te help standaardiseer. Dr. Lombard was ook betrokke by leksikografiese aktiwiteite buite die WAT, onder andere, as medewerker aan woordeboeke uitgegee deur OUPSA en gasdosent aan die Universiteit Stellenbosch. Hy het ook vier *Lexikos*-bydraes gelewer.

Dr. Lombard se toewyding aan die leksikografie, veral die praktyk, sowel as aan Afrikaans, is byna ongeëwenaard en sy heengaan is 'n enorme verlies vir ons vakgebied en vir die Afrikaanse taalkunde. Ons wil graag ons meegevoel aan elkeen van sy geliefdes, vriende, kollegas en medewerkers oordra.

Laastens wil ek graag my opregte waardering aan die AFRILEX-raad, die WAT en my mederedakteurs oordra vir hulle deurlopende vertroue en ondersteuning. Baie dankie ook aan AFRILEX vir die finansiële bystand wat die assosiasie aan *Lexikos* verleen. Ons is ook opreg dankbaar vir die goedgunstige en stewige finansiële bydrae van die Fonds Neerlandistiek.

André H. du Plessis
Hoofredakteur

Foreword

In 1989, a formal lexicographic discussion took place in Stellenbosch between the editorial staff of the *Woordeboek van die Afrikaanse Taal* (WAT — Dictionary of the Afrikaans Language) and a group of prominent linguists. These key figures realised that there existed a great need for the exchange of lexicographical information and experience in Africa. This ultimately led to the establishment of *Lexikos*, an excellent medium and vehicle for lexicographic discussion. *Lexikos* 1, a work of lexicographic pioneering, was published in 1991 under the editorship of the late Mr Pieter Hartevelt. For the first time, Africa had a journal that was primarily focused on lexicography. This journal has, 35 years later, undoubtedly become one of the most prominent journals for and about lexicography. The multitude of trendsetting *Lexikos* contributions over the last three and a half decades have had a lasting and significant impact on lexicography as a discipline. The editorial board, publisher (the WAT) and AFRILEX are immensely proud of the product that has been published over the past 35 years. May *Lexikos* benefit lexicography for the next 35 years and may many more volumes be successfully published!

As with the publication of *Lexikos* 1, the 35th volume also brought along challenges and tough decisions. Since 2022, the number of contributions submitted has systematically increased, but in 2025 the proverbial floodgates opened. The editorial team treaded water while trying to devise a plan to process and publish all the contributions. It was ultimately decided, with permission from the publisher, to publish two issues this year. This tremendous increase in contributions speaks to the reach and prestige of *Lexikos*. It is gratifying that so many authors consider *Lexikos* to be a suitable venue for their research outputs. However, this places unusual pressure on the editorial team, editorial support services and publisher. The increase in articles coupled with *Lexikos*' current editorial and financial model led to meaningful and necessary discussions between AFRILEX, the editorial team and the publisher about the sustainable management of the journal. The decision has been made that *Lexikos* 35 will be the last volume to charge page fees only to South African authors. From 2026 article processing charges will be implemented for authors outside South Africa. *Lexikos*' mission is to continue to ensure that knowledge is freely shared with the lexicographic community. Therefore, *Lexikos* remains an open access journal for all readers.

This year's first issue contains a special section dedicated to the 6th GLOBALEX Workshop on Lexicography and Neology. This workshop was presented with great success at the 28th International Conference of the African Association for Lexicography. Seven of the eleven papers were converted into articles and offer interesting and unique perspectives on the creation, use and lexicographic extraction of neologisms around the world. We are grateful for the collaboration with

GLOBALEX and appreciate the editorial and coordination work carried out by the guest editors of this section, Prof Annette Klosa-Kückelhaus and Mr Ilan Kernerman.

The other contributions in this year's issues are, as expected, representative of the variety of research avenues in lexicography and terminology. There are insightful and valuable contributions on, among others, terminology management, artificial intelligence, online dictionaries, metalexicography and inclusive lexicography. In addition to our authors from different parts of Africa, such as Botswana, Gabon and Tunisia, there are voices from, among others, Australia, China, Lithuania, Montenegro, Poland and Spain. A big thank-you to the authors who participate in the lexicographic discourse via their publications in *Lexikos*. The editorial team is eternally grateful for all the authors' cooperation and positive attitudes.

The publication of *Lexikos* is always a team effort, and the team that had a part in this year's issues is an extremely competent one. A word of thanks to Profs Elsabé Taljard, Dion Nkomo and Michele van der Merwe for the outstanding work delivered as associate editors, as well as for all their valuable guidance and assistance. Many thanks also to Dr Steve Ndinga-Koumba-Binza and Prof Sonja Bosch for their continued and extremely deft work as French and German editors. A sincere thank you to Dr Dané Claassen for the exceptional work with the reviews. It is a privilege to be able to rely on such a highly capable editorial team. On behalf of the editors, I would like to sincerely thank Mss Tanja Harteveld and Hermien van der Westhuizen from the WAT, for their devotion in promptly and professionally typesetting and finalising this year's issues, for their proficiency with the overall production of the journal, their administrative support and for their efficiency with all the ad hoc tasks that popped up during the year. I would also like to thank our newest team member, Ms Maryna Adshade. As editorial assistant, Ms Adshade has done a fantastic job and we are deeply grateful for her willingness, efficiency, dedication and support. A sincere thank you to Dr Phillip Louw, the Executive Director of the WAT, for his cooperativeness and commitment to *Lexikos*.

Lexikos is deeply indebted to each adjudicator that assisted with this year's peer review process, ensuring that our exacting scientific standards are maintained.

This year we learnt with heavy hearts about the passing of Dr Frikkie Lombard, former final editor of the WAT. Dr Lombard left a deep mark on the South African, especially Afrikaans, lexicography. Dr Lombard served the WAT with distinction for 42 years, and 25 years on the Taalkommissie (the South African Academy for Science and Arts' Language Commission). During his tenure, he made a lasting mark on the Afrikaans landscape, especially through the 10 WAT volumes on which he worked and the significant contribution he made to helping standardise the spelling rules of Afrikaans. Dr Lombard was also involved in lexicographic activities outside the WAT, among others, as a collaborator on dictionaries published by OUPSA and a guest lecturer at Stellenbosch University.

He also contributed four articles to *Lexikos*. Dr Lombard's dedication to lexicography, especially its practice, as well as to Afrikaans, is almost unparalleled and his passing is an enormous loss for our field and for Afrikaans linguistics. We would like to extend our condolences to each of his loved ones, friends, colleagues and associates.

Finally, I would like to express my sincere appreciation to the AFRILEX board, the WAT and my fellow editors for their continued trust and support. A big thank you to AFRILEX for the financial aid provided to *Lexikos*. We are also sincerely grateful for the kind and substantial financial contribution of the Fonds Neerlandistiek.

André H. du Plessis
Editor-in-Chief

'n Woord van AFRILEX

Na aanleiding van 'n goed georganiseerde 29ste kongres van die *African Association of Lexicography* (AFRILEX) wat by die Universiteit van Wes-Kaap gehou is, wens die uittredende AFRILEX-raad die *Lexikos*-span, onder leiding van die hoofredakteur, André du Plessis, en die uitgewer, die Buro van die WAT, geluk met die suksesvolle publikasie van hierdie bundel. Die Raad bly deurgaans bewus van die deeglike werk wat deur die *Lexikos*-redakteurs gedoen word. Hul volhardingsvermoë, stiptheid, noukeurigheid en deskundigheid in die veld sorg dat die hoë standaarde van die tydskrif gehandhaaf word.

Lexikos bly 'n belangrike mondstuk vir die toonaangewende navorsing van AFRILEX-lede, beide in Afrika en globaal. Die kongrestema vir die 29ste AFRILEX-kongres was "**Die rol van kunsmatige intelligensie en digitalisering in die toekoms van leksikografie**", wat baie interessante aanbiedings ontlok het wat die kruispunt van kunsmatige intelligensie en digitale gereedskap met die leksikografiese praktyk ondersoek het. Dit is 'n tydige en toekomsgerigte kongrestema aangesien kunsmatige intelligensie steeds in heelwat dissiplines oor die akademiese spektrum ontwrig en geleenthede bied.

Dit is ook noemenswaardig dat sommige van die navorsing wat by die 29ste AFRILEX-kongres aangebied is, gefokus het op navorsing wat in talle Suid-Afrikaanse tale gedoen is. Die toenemende aantal samewerkings oor instellings en fokusareas heen is 'n ware weerspieëling van 'n gedeelde doelwit om posities en met trefkrag by te dra tot die bevordering van ons veld. Hierdie inter- en kruisdissiplinêre aard kan slegs die diepte en breedte van die leksikografie as 'n dissipline verbeter. Ek hoop dat daardie interessante aanbiedings in artikels omskep word en in toekomstige volumes van *Lexikos* gepubliseer word.

Ek wil weereens die belangrike rol wat die Buro van die WAT oor die jare as uitgewer van *Lexikos* vertolk het, herhaal. Die Buro van die WAT bly 'n sleutel- en strategiese vennoot van AFRILEX. Namens die AFRILEX-Raad, en ook die algemene AFRILEX-lede wêreldwyd, wil ek die *Lexikos*-span, die Buro van die WAT en die outeurs wat bygedra het tot hierdie uitgawe, opreg bedank. Dit is nog 'n noemenswaardige bydrae tot die leksikografie as wetenskap.

Langa Khumalo
President: AFRILEX

A Few Words from AFRILEX

Emerging from a successfully organised 29th African Association of Lexicography (AFRILEX) conference held at the University of the Western Cape, the outgoing AFRILEX Board congratulates the *Lexikos* team, under the stewardship of the Editor-in-Chief, André du Plessis, and the publisher, the Bureau of the WAT, for the successful publication of this volume. The Board remains mindful of the sterling work that is done by the *Lexikos* editors. Their tenacity, stringency, meticulous attention to detail, and expertise in the field has helped maintain the very high standards of the journal.

Lexikos remains an important academic mouthpiece of the cutting-edge research that is done by the AFRILEX members both in the continent and globally. The conference theme for the 29th AFRILEX conference was **'The Role of Artificial Intelligence and Digitalisation in the Future of Lexicography'**, which drew very interesting presentations that explored the intersection of artificial intelligence and digital tools with lexicographic praxis. The conference theme was timely and forward-looking, as artificial intelligence continues to disrupt and present opportunities in many disciplines across the academic spectrum.

It is also notable that some of the research presented at the 29th AFRILEX conference focused on research done in numerous South African languages. The increased number of collaborations across institutions and focus areas is a true reflection of a shared goal to contribute positively and impactfully to the advancement of our field. This inter and cross-disciplinarity can only enhance the depth and breadth of lexicography as a discipline. I hope to see those interesting presentations developed into articles and published in future volumes of *Lexikos*.

I want to reiterate the important role that the Bureau of the WAT has played over the years as the publisher of *Lexikos*. The Bureau of the WAT remains a key and strategic partner of AFRILEX. On behalf of the AFRILEX Board, and indeed the general members of AFRILEX world-wide, I sincerely thank the *Lexikos* team, the Bureau of the WAT, and the contributing authors to this *Lexikos* volume. It is another important contribution to the scholarship of lexicography.

Langa Khumalo
President: AFRILEX

Redaksionele doelstellings

Lexikos is 'n tydskrif vir die leksikografiese vakspesialis en word in die AFRILEX-reeks uitgegee. "AFRILEX" is 'n akroniem vir "leksikografie in en vir Afrika". Van die sesde uitgawe af dien *Lexikos* as die amptelike mondstuk van die *African Association for Lexicography* (AFRILEX), onder meer omdat die Buro van die WAT juis die uitgesproke doel met die uitgee van die AFRILEX-reeks gehad het om die stigting van so 'n leksikografiese vereniging vir Afrika te bevorder.

Die strewe van die AFRILEX-reeks is:

- (1) om 'n kommunikasiekanaal vir die nasionale en internasionale leksikografiese gesprek te skep, en in die besonder die leksikografie in Afrika met sy ryk taleverskeidenheid te dien;
- (2) om die gesprek tussen leksikograwe onderling en tussen leksikograwe en taalkundiges te stimuleer;
- (3) om kontak met plaaslike en buitelandse leksikografiese projekte te bewerkstellig en te bevorder;
- (4) om die interdisiplinêre aard van die leksikografie, wat ook terreine soos die taalkunde, algemene taalwetenskap, leksikologie, rekenaarwetenskap, bestuurskunde, e.d. betrek, onder die algemene aandag te bring;
- (5) om beter samewerking op alle terreine van die leksikografie moontlik te maak en te koördineer, en
- (6) om die doelstellings van die *African Association for Lexicography* (AFRILEX) te bevorder.

Hierdie strewe van die AFRILEX-reeks sal deur die volgende gedien word:

- (1) Bydraes tot die leksikografiese gesprek word in die vaktydskrif *Lexikos* in die AFRILEX-reeks gepubliseer.
- (2) Monografiese en ander studies op hierdie terrein verskyn as afsonderlike publikasies in die AFRILEX-reeks.
- (3) Slegs bydraes wat streng vakgerig is en wat oor die suiwer leksikografie of die raakvlak tussen die leksikografie en ander verwante terreine handel, sal vir opname in die AFRILEX-reeks kwalifiseer.
- (4) Die wetenskaplike standaard van die bydraes sal gewaarborg word deur hulle aan 'n komitee van vakspesialiste van hoë akademiese aansien voor te lê vir anonieme keuring.

Lexikos sal jaarliks verskyn, terwyl verdienstelike monografiese studies sporadies en onder hulle eie titels in die AFRILEX-reeks uitgegee sal word.

Editorial Objectives

Lexikos is a journal for the lexicographic specialist and is published in the AFRILEX Series. "AFRILEX" is an acronym for "lexicography in and for Africa". From the sixth issue, *Lexikos* serves as the official mouthpiece of the *African Association for Lexicography* (AFRILEX), amongst other reasons because the Bureau of the WAT had the express aim of promoting the establishment of such a lexicographic association for Africa with the publication of the AFRILEX Series.

The objectives of the AFRILEX Series are:

- (1) to create a vehicle for national and international discussion of lexicography, and in particular to serve lexicography in Africa with its rich variety of languages;
- (2) to stimulate discourse between lexicographers as well as between lexicographers and linguists;
- (3) to establish and promote contact with local and foreign lexicographic projects;
- (4) to focus general attention on the interdisciplinary nature of lexicography, which also involves fields such as linguistics, general linguistics, lexicology, computer science, management, etc.;
- (5) to further and coordinate cooperation in all fields of lexicography; and
- (6) to promote the aims of the *African Association for Lexicography* (AFRILEX).

These objectives of the AFRILEX Series will be served by the following:

- (1) Contributions to the lexicographic discussion will be published in the specialist journal *Lexikos* in the AFRILEX Series.
- (2) Monographic and other studies in this field will appear as separate publications in the AFRILEX Series.
- (3) Only subject-related contributions will qualify for publication in the AFRILEX Series. They can deal with pure lexicography or with the intersection between lexicography and other related fields.
- (4) Contributions are judged anonymously by a panel of highly-rated experts to guarantee their academic standard.

Lexikos will be published annually, but meritorious monographic studies will appear as separate publications in the AFRILEX Series.

Die gebruik van kleur in aanlyn woordeboeke, met oorweging van drie Afrikaanse woordeboeke

Maryna Adshade, *Buro van die WAT, Stellenbosch,
Suid-Afrika; Departement Afrikaans en Nederlands,
Universiteit Stellenbosch, Stellenbosch, Suid-Afrika*
(adshadem@sun.ac.za)

(<https://orcid.org/0009-0002-1491-7488>)

Opsomming: Wanneer die gebruik van kleur in sommige aanlyn Afrikaanse woordeboeke oorweeg word, wil dit voorkom asof kleur soms oneffektief of lukraak ingespan word. In ander gevalle kom dit voor asof kleure gekies word om vinnig met die handelsmerkidentiteit van die woordeboekuitgewer geassosieer te word eerder as ter wille van praktiese doeleindes. Dit kan moontlik weens abstrakte of vae voorstelle vanuit die metaleksikografie wees wat onbevredigende leiding aan praktiese leksikograwe bied, maar dalk ook 'n gevolg van die Suid-Afrikaanse, en Afrikaanse, leksikografie wat relatief stadig tot die Westerse wêreld ontwikkel. Die gebruik van kleur word in hierdie artikel bespreek met oorweging van 'n paar voorstelle wat reeds in metaleksikografiese navorsing gelug is. Hierdie voorstelle word aangevul met enkele riglyne ontleen aan die inligtingstechnologie. Die toepassing, al dan nie, van die teoretiese moontlikhede word bespreek ten opsigte van die aanlyn weergawe van die *Handwoordeboek van die Afrikaanse Taal*, *Aanlyn Woordeboek van die Afrikaanse Taal* en die *Woordeboek van Afrikaans Vandag*. Daar is ruimte vir verbetering in sekere aspekte van hierdie woordeboeke se kleurgebruik, en heelwat van die opwindende, meer gevorderde moontlikhede wat deur 'n beter kleurgebruik ingehou word, is nog toekomsmusiek binne die Afrikaanse leksikografie. Nietemin is dit opmerklik dat die gebruik van kleur belowend is en vorentoe 'n groter rol in toeganklikheid kan speel.

Sleutelwoorde: KLEURGEBRUIK, AANLYN WOORDEBOEKE, DATATOEGANKLIKHEID, WOORDEBOEKSTRUKTURE, ARTIKELSTRUKTUUR

Abstract: The Use of Colour in Online Dictionaries, with Consideration of Three Afrikaans Dictionaries. When considering the use of colour in some online Afrikaans dictionaries, it seems that colour is sometimes used ineffectively or haphazardly. In other cases, colours seem to be chosen mainly to align with the dictionary publisher's brand identity, rather than for practical purposes. This may result from abstract or vague metalexigraphical suggestions that offer unsatisfactory guidance to practical lexicographers but could also be the result of South African, and Afrikaans, lexicography developing slower compared to the Western world. This article discusses the use of colour in online dictionaries, considering a few suggestions from metalexigraphical research. These suggestions are then supplemented with some guidelines derived from

the field of information technology. The application of the theoretical possibilities, or lack thereof, is discussed with regard to the online version of the *Handwoordeboek van die Afrikaans Taal*, *Aanlyn Woordeboek van die Afrikaans Taal* and the *Woordeboek van Afrikaans Vandag*. These dictionaries show room for improvement in some aspects related to colour usage, with many exciting and more advanced possibilities still a future prospect for Afrikaans lexicography. Nevertheless, it is noteworthy that the use of colour is promising and may play a greater role in accessibility in the future.

Keywords: COLOUR USAGE, ONLINE DICTIONARIES, DATA ACCESSIBILITY, DICTIONARY STRUCTURES, ARTICLE STRUCTURE

1. Inleiding

Dit is algemeen bekend dat die leksikografiepraktyk deur die ontwikkeling en geskiedenis daarvan reeds verskeie uitdagings en paradigmaterskuiwings moes deurstaan (sien Gouws 2017: 430-438). Dit is ook goed bekend dat dit tans opnuut deur 'n enorme paradigmaterskuiwing uitgedaag word, naamlik die oorgang van die gedrukte na die elektroniese medium, veral na die aanlyn sfeer. In die leksikografiepraktyk se geskiedenis is hierdie oorgang onbetwisbaar die ingrypendste oorgang nóg (Gouws 2018c: 178) en dit gaan sodoende uiteraard met nuwe uitdagings gepaard. Dit is teweeggebring deur die koms van die era van elektroniese dataverwerking, en kan sedertdien toegeskryf word aan 'n bestendige stroom tegnologiese deurbreke wat gelei het tot radikaal nuwe benaderings in onder meer hoe leksikografiese produkte gefinansier, voorberei, saamgestel, gebruik en aangebied word (Gouws en Tarp 2024: 190).

Dit is juis op 'n aspek van die laasgenoemde vlak van vernuwing waarop hierdie artikel toegespits is. Soos Almind (2005: 37) aandui, is daar ongelukkig uitgewers (en dalk selfs leksikograwe) wat glo dat hulle 'n behoorlike aanlyn woordeboek kan publiseer deur gewoon 'n gedrukte woordeboek te digitiseer en 'n soekenjin op die gebruikerskoppelvlak van daardie digitale weergawe aan te bring. Hy (ibid.) meen dat dié uitgewers verkeerdelik veronderstel dat potensiële woordeboekgebruikers daarmee tevrede sal wees om 'n gedrukte ontwerp op 'n skerm te sien vertoon. Dit lei tot twee probleme: Daar word nie daarmee tred gehou dat 'n aanlyn woordeboek 'n produk in eie reg is nie, en as bemarkingstrategie sal die benadering misluk bloot omdat dit woordeboekgebruikers sal teleurstel (Almind 2005: 37). Hy is daarom oortuig daarvan dat leksikograwe by die beplanning van aanlyn woordeboeke dit eerder van meet af as sodanig moet beplan en nie slegs 'n gedrukte woordeboek aanlyn beskikbaar moet stel nie (Almind 2005: 37). Die grootste rede hiervoor kan toegeskryf word aan een kernbegrip: toeganklikheid (Almind 2005: 38).

Hoewel reeds twee dekades gelede gelug, geld dié siening dat baie aanlyn woordeboeke gewoon gedigitiseerde gedrukte woordeboeke is, vandag steeds heelwat aanlyn woordeboeke. Boonop bly dit, ondanks meer onlangse vordering, besonders relevant in die Suid-Afrikaanse (en Afrikaanse) leksikografielandskap, waar die ontwikkeling op vele terreine steeds stadiger as verwag vorder

en die aanlyn leksikografie veral agter dié van Westerse lande bly (Beyer en Louw 2022: 41). Met inagneming hiervan het hierdie artikel ten doel om een ontwerpelement van aanlyn woordeboeke, naamlik die gebruik van kleur, te bespreek. Die doel is ook om te wys op 'n paar opwindende moontlikhede en om enkele voorstelle omtrent suksesvolle aanwending van kleur te maak.

Om die bespreking hiervan te kontekstualiseer, moet die gebruik van kleur eers kortliks relatief tot die voorkoms van aanlyn woordeboeke en woordeboekstruktuur oorweeg word. Ná die teoretiese bespreking word daar gekyk hoe kleur in drie Afrikaanse woordeboeke aangewend word, naamlik die aanlyn weergawe van die sesde uitgawe van die *Handwoordeboek van die Afrikaanse Taal* (voortaan HAT Aanlyn), die *Aanlyn Woordeboek van die Afrikaanse Taal* (voortaan Aanlyn WAT)¹, en die *Woordeboek van Afrikaans Vandag* (voortaan WAV).

2. Teoretiese beskouinge oor die gebruik van kleur in woordeboeke

Voordat die gebruik van kleur in woordeboeke bespreek kan word, is dit nodig om daarop te wys dat kleurgebruik nie slegs 'n kosmetiese aanvulling tot die data-aanbieding in woordeboeke behoort te wees nie. Kleur moet eerder as nuttige hulpmiddel in 'n woordeboek se strukturele aanbod ingespan word om te verseker dat aanlyn woordeboeke as funksionele en praktiese gebruiksinstrumente kan dien. Kleur moet dus beskou word as 'n praktiese middel tot 'n doel en 'n funksionele element in woordeboekartikels.

2.1 Die verband tussen woordeboekstrukture en kleurgebruik

Woordeboeke, ook aanlyn woordeboeke, kan beskou word as draers van tekssoorte, oftewel groot tekste opgebou uit kleiner, gerangskikte tekste (Wiegand 1996: 135-136). Hierdie tekste moet só aangebied word dat die struktuur daarvan die doel van die tersaaklike tekste kan vervul (Wiegand 1983: 104). Dit beteken dat dit só gerangskik moet word dat woordeboeke as praktiese gebruiksinstrumente kan dien om woordeboekgebruikers se leksikografiese behoeftes te bevredig (Wiegand 1998: VII, 62).

Vir woordeboeke om nutsprodukte te wees, moet daar volgens Gouws (2018b: 217-218) by die beplanning van enige woordeboek, ongeag die medium, met drie integrale komponente tred gehou word: leksikografiese funksie(s), woordeboekinhoud en woordeboekstrukture. Die behoeftes en naslaanvaardighede van die beplande woordeboek se beoogde teikengebruikers behoort eerstens die funksie(s) te bepaal, waarna die inhoud op grond van die funksie(s) vasgestel moet word (sien Gouws 2018b: 217-218). Uiteindelik moet die inhoud op 'n gepaste wyse aangebied word deur die inspanning van verskillende relevante woordeboekstrukture (Gouws 2018b: 217). Gepaste woordeboekstrukture moet dus in oorleg met funksies en inhoud bepaal word.

Uit hierdie komponente is die eerste twee redelik geyk in die metaleksikografie, en in 'n hoë mate tans nog vanselfsprekend omdat dit grotendeels onver-

anderd in aanlyn woordeboeke geld. Die derde komponent kan egter te midde van die huidige paradigmaterskuiwing onsekerheid en probleme lewer, want gedrukte woordeboekstrukture kan nie altyd min of meer onveranderd in aanlyn woordeboeke behou word nie (Gouws 2018b: 218). 'n Besinning hieroor, en die faktore wat dit kan beïnvloed, is daarom nodig omdat dié strukture voorsiening maak vir leksikografiese data (sowel binne of buite woordeboekartikels) sodat woordeboekgebruikers data kan opspoor (Gouws 2023: 383). Hieruit is woordeboekstrukture duidelik onontbeerlik — daarsonder sou woordeboeke, ongeag die gehalte van die woordeboekinhoud, onbevredigende produkte wees omdat woordeboekgebruikers nie vinnig en ondubbelsinnig hulle verlangde inligting daaraan sou kon onttrek nie (De Schryver 2003: 173).

Hier is die fokus nie op 'n omvattende bespreking van woordeboekstrukture nie. Daar word ook aanvaar dat heelwat van die hieropvolgende terminologie algemeen gangbaar is sodat alles nie in diepte bespreek word nie. Om die rol en die aanbod van kleurgebruik in woordeboekstrukture te verstaan, is dit nietemin nuttig om kortliks na kernstrukture en gidsstrukture te verwys, asook en veral na die interne soekstruktuur as tipe gidsstruktuur.

Alhoewel die onderskeid nie 'n verabsoluttering is nie en oorvleuelings tussen en binne die kategorieë kan bestaan, kan kernstrukture as die hoofstrukture van woordeboeke beskou word terwyl gidsstrukture die sekondêre vlak uitmaak. Kernstrukture hou verband met die beskikbaarheid en posisionering van leksikografiese data (Mihindou 2006: 163) en sluit onder meer die makrostruktuur, mikrostruktuur en dataverspreidingstruktuur in (vgl. Gouws 2005: 63). Hierteenoor geld gidsstrukture byvoorbeeld die toegangstruktuur, mediostruktuur en sekere buitetekste (vgl. Gouws 2001: 104; Gouws 2005: 63), oftewel daardie strukture wat woordeboekgebruikers help om data op te spoor en van die woordeboekraadplegingsproses sin te maak (Mihindou 2006: 163).

Dit is spesifiek by gidsstrukture, soos die toegangstruktuur, waar kleur in aanlyn woordeboeke 'n groter rol kan en moet speel, en waar leksikograwe doelgerig te werk moet gaan om dit te bereik. Die toegangstruktuur bied die soekroete wat woordeboekgebruikers volg wanneer hulle woordeboeke raadpleeg om by die verlangde woordeboekartikel uit te kom (Gouws en Prinsloo 2005: 165). Dit funksioneer in samewerking met ander strukture — veral die dataverspreidingstruktuur, mikrostruktuur, mediostruktuur en artikelstruktuur — en het die belangrike doel om woordeboekgebruikers met so min as moontlik moeite die benodigde data te laat bereik (Gouws en Prinsloo 2005: 172). Daar kan onderskei word tussen 'n eksterne en interne soekstruktuur, maar daar word vervolgens slegs op die laasgenoemde gefokus. Vir meer oor die eksterne soekstruktuur kan Gouws en Prinsloo (2005), asook Wiegand en Beer (2013) geraadpleeg word.

By die interne soekstruktuur gaan dit oor die soekroete wat woordeboekgebruikers binne woordeboekartikels volg om die verlangde datatipe of datakategorie te bereik (Gouws en Prinsloo 2005: 170) sodat hulle aan hulle onmiddellike inligtingsbehoefte kan voldoen (vgl. Tarp 2012: 328). Soos Wiegand en Beer (2013: 139) dit stel, ondersteun die interne toegangstruktuur woordeboekgebruikers om hulleself binne artikels te oriënteer nadat hulle eksterne tekstuele toegangs-

elemente gevolg het om die tersaaklike artikel te bereik.

Hier is dit nodig om weer te verwys na De Schryver (2003: 173) se punt dat woordeboeke onbevredigende produkte sal wees as woordeboekgebruikers nie vinnige, ondubbelsinnige toegang tot data kan verkry nie. Gouws en Prinsloo (2005: 170, 172) se opmerking moet hiermee in verband gebring word wanneer hulle tereg noem dat woordeboekgebruikers selde die hele woordeboekartikel wil lees om die verlangde inligting te onttrek, en daarom meer waardering het vir woordeboeke waarin hulle vinniger toegang tot die verlangde data kan kry. Hieruit is dit duidelik dat die interne toegangstruktuur noodsaaklik is om vinnige toegang tot relevante data te verseker, en soos Wiegand en Beer (2013: 139) aandui, des te meer in langer gekondenseerde woordeboekartikels. Dit is ook hier waar kleur, indien dit effektief en funksioneel ingespan word, 'n waardevolle rol in vinniger, ondubbelsinnige toegang kan speel.

2.2 Moontlikhede vir die gebruik van kleur as toegangselement

Die vinniger en ondubbelsinnige toegang tot data in woordeboekartikels hou naas die posisionering van data verband met hoe data aangebied word, oftewel die voorkoms van 'n woordeboekartikel. Die leksikografie is nie geïsoleerd nie, soos duidelik blyk uit die oorweging van tegnologie en die ontwikkeling daarvan in die huidige paradigmaterskuiwing, soos genoem in die inleiding. In die vinnig ontwikkelende digitale wêreld waar mense se digitale geletterdheid toenemend verbeter, is dit daarom 'n wesenlike kwessie dat woordeboeke ontwikkel moet word om steeds relevant en bruikbaar te wees, ook ten opsigte van hoe dit visueel by ander eietydse digitale bronne aansluit.

Soos Gouws (2018c: 179-180) aandui, is dit nodig om te besef dat heelwat potensieële toekomstige woordeboekgebruikers tot nuwe geslagte, soos Generasie Z, behoort en hierdie geslagte volwaardige digitale burgers is vir wie die internet 'n gegewe is. Die vanselfsprekendheid van die internet in hulle lewe, asook in dié van baie lede van ander generasies, beïnvloed hoe hulle met digitale hulpbronne werk en wat hulle verwagtinge daarvan (en van die voorkoms daarvan) is. Dit is dus vir lewensvatbaarheidsredes noodsaaklik om tred te hou met gebruikers se vorderende of gevorderde digitale geletterdheid, asook hulle verwagtinge van digitale bronne. In die huidige digitale era waartydens digitale burgers dikwels bederf is met vinnige, moeitelose toegang tot 'n oorfloed inligting (vgl. Gouws en Tarp 2017: 394), sal woordeboeke wat outyds lyk of grotendeels 'n gedrukte woordeboek op 'n skerm is, onbevredigend wees en dalk in onbruik verval. Met die merkwaardige tempo waarteen kunsmatige intelligensie ontwikkel en die gebruikers daarvan van pasgemaakte oplossings kan verskaf, is dit nou dalk meer toepaslik as ooit vantevore.

Aanlyn woordeboeke moet daarom, soos reeds genoem, as 'n produk in eie reg beskou word en meer wees as 'n elektronies beskikbare gedrukte woordeboek (vgl. Almind 2005: 37). As aanlyn produkte kan daar tekens van vernuwing en ooreenkomste met bepaalde internetpraktyke wees, ook ten opsigte

van die ontwerp en algemene voorkoms van woordeboeke. Rundell (2015: 320) stel byvoorbeeld dat leksikograwe kan aanvaar dat woordeboekgebruikers vertrouwd is met internetbeginsels, soos soekstrategieë en 'n sekere "woordeskat", maar dit geld myns insiens ook vaardighede om internetbronne te kan navigeer, waarin kleur 'n rol kan speel. Hierdie veronderstelling word bevestig deur Müller-Spitzer, Koplenig en Töpel (2012: 441) wat daarop wys dat woordeboekgebruikers 'n behoefte het aan aanlyn woordeboeke wat bekendheid toon met ander toepassings en webformateringsbeginsels.

Een vorm hiervan is die gebruik van kleur volgens bepaalde formateringsbeginsels. Swasty en Adriyanto (2017: 17-18) wys byvoorbeeld vyf ontwerpbegin-sels van goeie gebruikerskoppelvlakke uit, waarvan een die goeie gebruik van kleur is. Kleur kan 'n aanwys in gebruikerskoppelvlakke se ontwerp wees (Jamil en Denes 2024: 53) deurdat dit 'n wesenlike rol in estetiese waarde, bruikbaarheid en toeganklikheid kan speel (WCAG 2024). Alhoewel die hoofdoel van aanlyn woordeboeke uiteraard nie is om estetiese waarde te hê nie, bly estetika 'n faktor omdat gebruikers se persepsie van gebruikerskoppelvlakke se estetiese waarde hulle persepsie van die bruikbaarheid daarvan kan beïnvloed. In hierdie verband verwys Alharoon en Gillan (2020: 1876) na studies wat 'n positiewe korrelasie toon tussen skynbare en inherente bruikbaarheid, dit wil sê hoe gebruikers 'n koppelvlak se gebruiksgemak skat soos gebaseer op voorkoms (nog voordat hulle dit gebruik het), teenoor die vasstelling van gebruiksgemak nadat hulle met die koppelvlak gewerk het. Dus kan aanlyn woordeboeke wat onafgerond, oudmodies of andersins onesteties is gebruikers ontmoedig om die woordeboek te gebruik. Dit sluit weer aan by Almind (2005: 37) se siening dat 'n slaafse navolging van gedruktewoordeboekbeginsels as bemarkingstrategie kan misluk omdat woordeboekgebruikers teleurgesteld kan wees in die produk.

Die kwessie van estetika en kleur as 'n belangrike ontwerpbegin-sel is dus ook op woordeboekontwerp van toepassing. Soos Mihindou (2006: 165) aantoon, verkies woordeboekgebruikers daardie woordeboeke waarin kleur gebruik word omdat kleur aantrekliker en meer ekspressief is. As 'n aanwys behoort kleurgebruik egter nie net kosmetiese waarde te hê nie en dit moet funksioneel wees. Hier is Müller-Spitzer en Koplenig (2014: 144) se siening dat vorm die funksie in 'n woordeboek moet volg, relevant. Hierbenewens moet kleur in same-werking met ander elemente gebruik word en kleur kan nie die enigste manier wees waarop sekere inligting oorgedra word nie.

In die interne toegangstruktuur kan verskillende kleure woordeboekgebruikers toelaat om vinniger en makliker toegang tot die verlangde data te verkry (Mihindou 2006: 165). Eweneens wys Kilgarriff en Kosem (2012: 49) daarop dat kleurgebruik navigasie kan vergemaklik deur woordeboekgebruikers te help om tussen verskillende datatipes te onderskei. Hulle noem nie spesifieke voorbeelde nie, maar verskillende kleure kan byvoorbeeld as struktuurmerkers gebruik word om soeksones in aanlyn woordeboeke te ekspliseer en duidelik van ander af te baken. Kleur kan sodoende tot 'n meer gebruikersvriendelike mikro-argi-tektuur bydra deurdat sekere elemente visueel van mekaar onderskei en in spesifieke kleure geblok word. Kundige woordeboekgebruikers met 'n goeie woor-

deboekcultuur kan sodoende vinniger data opspoor deur die oog gewoon na die verlangde deel te laat gly. Die voorwaarde is egter dat 'n verduideliking van die betekenis van spesifieke kleure in die gebruikersgids, 'n sleutel onderaan die artikels, of iets dergeliks ingesluit moet word.

Benewens die skeiding en duidelike afbakening van verskillende soeksones, kan kleur woordeboekgebruikers help om op data af te kom wat hulle andersins dalk nie (so maklik) sou vind nie. As woordeboekgebruikers met 'n woordeboek se benadering tot kleur bekend is, kan die gebruik van verskillende kleure hulle aandag trek na soeksones wat in eenkleurige artikels tussen ander soeksones sou kon verdwyn (vgl. Mihindou 2006: 165). Hierbenewens kan kleur gebruik word om mediostrukturele verwysings duideliker sigbaar te maak en van omliggende data te onderskei, wat ook gebruikers na meer data as net die onmiddellik verlangde data kan lei. Soos tipies in internetbronne, kan mediostrukturele verwysings wat hiperskakels is as sodanig met kleur aangedui word sodat woordeboekgebruikers met 'n oogopslag kan sien waar kits-toegang tot ander verwante artikels en data moontlik is. Dit kan veral by implisiete kruisverwysings nuttig wees wat andersins maklik misgekyk kan word. Verder kan verskillende kleure gebruik word om woordeboekinterne en woordeboek-eksterne kruisverwysings te onderskei. Hierdie opsies kan in verskillende mates tot politoeganklike en meer gebruikersvriendelike woordeboekartikels bydra.

Kleur kan dus gebruik word om verbande tussen verskillende dele van 'n woordeboek te lê. Mihindou (2006: 166) is byvoorbeeld ten gunste van die gebruik van bypassende kleure sodat woordeboekgebruikers verskillende tekste binne die woordeboek kan identifiseer — spesifiek deur die sentrale teks van buitetekste te onderskei. Kleur kan só as 'n tipe indeks dien omdat gebruikers vinnig kan sien waar verskillende dele van die woordeboek begin. Hierdie voorstel is uiteraard minder ter sake in aanlyn woordeboeke waarin hiperskakelroetes na verskillende woordeboekdele (wat dikwels op 'n navigasiebalk vertoon word) moontlik is. 'n Aanpassing van Mihindou se voorstel kan wel oorweeg word: Bypassende kleure kan binne woordeboekartikels en buitetekste gebruik word om woordeboekgebruikers bewus te maak van waar hulle addisionele data kan vind wat ook nuttig kan wees. Indien woordeboekgebruikers binne 'n bepaalde artikel op 'n spesifieke kleur afkom, kan dit hulle byvoorbeeld na geïntegreerde buitetekste verwys, oftewel daardie buitetekste wat tot 'n woordeboek se werklike doel bydra (Klosa en Gouws 2015: 144). Andersins kan dit die verhouding tussen dele van die sentrale teks en spesifieke buitetekste, of spesifieke dele van bepaalde buitetekste, aandui wanneer byvoorbeeld sekere struktuurmerkers, teksdele, rame om teksdele, of die skermkootagtergrond in spesifieke kleure gemerk word.

In meer gevorderde aanlyn woordeboeke kan kleur ook in artikels met 'n dinamiese veelvlaklike artikelstruktuur 'n waardevolle rol aanneem. Dié artikels kan omvattend of beperk wees: Omvattende artikels is volle artikels waarvan die geheelbeeld nie in 'n enkele skermkoot sigbaar is nie terwyl beperkte artikels dele van omvattende artikels is wat in aparte skermkote verskyn (Gouws 2018a: 52). Binne veelvlaklike artikels kan 'n soektonnel voorkom wat woorde-

boekgebruikers toelaat om tot die beperkte artikels toegang te verkry. Gouws en Bothma (2024: 362) beskryf 'n soektonnel as 'n struktuur waardeur 'n gebruiker stap vir stap deur verskillende artikelvlakke (die beperkte artikels) kan beweeg. Hulle beweeg dan horisontaal of sywaarts deur 'n vlak om verdere kommentaar of subkommentaar binne bepaalde vakke te bereik (Gouws 2023: 386; Gouws en Bothma 2024: 362). Hier sou 'n vertikale beweging tussen verskillende vlakke, via 'n struktuur soos 'n soekskag, ook moontlik wees.

Om deur soektonnels, of ook soekskagte, te beweeg, moet gebruikers eers op die behoefte aan addisionele data dui deur byvoorbeeld op knoppies of ander struktuurmerkers te klik (Gouws 2018a: 50). Vir hierdie beweging, hetsy na verskillende vakke of vlakke, sal 'n meer komplekse koppelvlakontwerp egter benodig word (Gouws en Bothma 2024: 361). Dit is juis by hierdie aanduiding van gebruikers se behoefte aan nog data en in meer komplekse koppelvlakontwerpe waar kleur nuttig kan wees. Dit is omdat bepaalde kleure gebruik kan word om daardie knoppies of struktuurmerkers duideliker te maak. So kan kleur dit duideliker maak waar gebruikers op versoek beperkte artikels kan vertoon, oftewel waar addisionele data beskikbaar is indien die gebruiker dit sou wou sien. Hier kan kleur ook indirek 'n rol speel in gebruikers se gemak en vermoë om verrykte en uitgebreide data te bekom.

By 'n veelvlakkige artikelstruktuur, soektonnels en soekskagte kan kleur nogmaals 'n rol speel om verbande tussen verskillende dele van 'n woordeboek te illustreer. Die addisionele data wat via sodanige soektonnels of soekskagte bekom kan word, word in 'n nuwe skermskoot vertoon wat aan die tersaaklike data-tipe (soos voorbeeldmateriaal, verwante leksikale items, grammatika of illustrasies) toegewy is (Gouws 2018a: 50). Hier kan kleur veral in beperkte artikels binne soektonnels help om byvoorbeeld dele van die sentrale teks aan ander relevante dele te merk, of die sentrale teks aan spesifieke geïntegreerde buitetekste te verbind, soos bo genoem. Daarbenewens kan dit 'n rol in niegeïntegreerde buitetekste speel, dit wil sê daardie buitetekste waaraan woordeboekgebruikers nie data kan onttrek wat tot 'n woordeboek se doel bydra nie (Gouws en Prinsloo 2005: 59), soos 'n woordeboek se beleid of gebruikersgids. In laasgenoemde kan kleur byvoorbeeld wys hoe en waar spesifieke kleure in artikels voorkom om aanduiders en struktuurmerkers te merk. Tydens hulle woordeboekraadpleging kan dit woordeboekgebruikers weer help om struktuurmerkers te sien wat hulle byvoorbeeld via 'n hiperskakelroete of opskietvenster vanuit 'n woordeboek-artikel kan neem na 'n niegeïntegreerde buiteteks of 'n gedeelte daarvan, soos 'n hanteringsbeleid.

2.3 Oorwegingspunte ten opsigte van doeltreffende kleurgebruik

Ofskoon kleur op verskeie maniere geïntegreer kan word, moet dit nie op 'n lukrake wyse geskied en bloot ingesluit word omdat die aanlyn wêreld die opsie bied nie. Dit beteken eerstens dat kleur nie oordadig gebruik moet word nie. Volgens Almind (2005: 39) moet kleur slegs gebruik word om belangrike data

uit te lig, en kleurgebruik moet tot "goeie" (kontrasterende) kleure beperk word. Andersins kan oormatige kleurgebruik vinnig tot 'n sirkusontwerp lei (Almind 2005: 52), wat beteken dat dit byvoorbeeld nes die oorgebruik van tipografiese struktuurmerkers, soos kursief, vinnig steurend kan word. Dit kan lei tot 'n onaantreklike en deurmekaar voorkoms, maar kan ook teksverdigting veroorsaak wat tot 'n stadiger, potensieel dubbelsinnige inligtingonttrekking lei. Die uitdaging by die beplanning van die gebruikerskoppelvlak is dus die fyn balans tussen te veel en te min kleur.

Met verwysing na Almind (2005) se eise aan aanlyn woordeboeke, merk Fuertes-Olivera en Tarp (2014: 196-197) op dat kleur tot 'n elegante, aangename uitleg asook tot woordeboekartikels se leesbaarheid moet bydra deurdat maklik leesbare en kontrasterende kleure gekies moet word. Deur laasgenoemde toe te pas en elke datatipe in 'n nuwe reël te laat begin, kan die leesbaarheid van woordeboekartikels verhoog word (Fuertes-Olivera en Tarp 2014: 197). Dit kan ook 'n gebruikersvriendelike mikro-argitektuur bevorder deur woordeboekgebruikers toe te laat om vinnig by die verlangde data uit te kom.

Betreffende 'n elegante, aangename uitleg kan die keuse van kleure gebaseer wees op die gebruikerskoppelvlakontwerpers se persepsie van estetiese waarde (Jamil en Denes 2024: 54). Dit kan andersins gebaseer word op kleure wat pas by die tersaaklike woordeboek se kleurskema en dit wat daarmee geassosieer word (vgl. Almind 2005: 51). 'n Beperkte kleurpalet as verkose benadering kan om kosmetiese redes aanloklik wees, onder andere as dit pas by 'n woordeboek se beeld en die kleure wat daarmee geassosieer word. In die WAT se geval is die kleuridentiteit byvoorbeeld kenmerkend blou, soos prominent op die WAT se korporatiewe blad (<https://www.wat.co.za/>). 'n Woordeboek se kleuridentiteit kan vir handelsmerkbestuurdoeleindes belangrik wees (vgl. Swasty en Adriyanto 2017: 22), maar by produkte soos woordeboeke waarvan die primêre doel praktiese bruikbaarheid is, moet die funksionaliteit van kleurgebruik ook oorweeg word.

Ten spyte van die moontlike estetiese voordeel, kan 'n beperkte kleurpalet vir sommige woordeboekgebruikers nie funksioneel wees. Dit geld byvoorbeeld woordeboekgebruikers wat sukkel om tussen kleure te onderskei, byvoorbeeld vanweë kleurblindheid, vir wie skakeringe van hoofsaaklik een kleur (anders as swart of wit) of spesifieke kleurkombinasies nie funksioneel of moontlik selfs verwarrend of hinderlik kan wees. Die bedoelde betekenis van die kleure as tipe struktuurmerkers gaan dus verlore — bowenal in die afwesigheid van 'n sleutel of kort verduideliking van die kleure se betekenis op die artikelskermskoot. Daarbenewens kan lae kontrasterende kleure tot 'n aansienlik verminderde estetiese ervaring lei vir gebruikers wat nie so maklik tussen kleure differensieer nie (Jamil en Denes 2024: 54). Dus kan 'n beperkte kleurgebruik vir hierdie woordeboekgebruikers nóg funksioneel, nóg esteties wees.

Die gebruik van kontrasterende kleure is derhalwe belangrik. Nietemin gee nie Almind (2005) of Fuertes-Olivera en Tarp (2014) 'n bevredigende, konkrete verduideliking van wat in woordeboeke relatief tot wat moet kontrasteer nie. Buiten die tradisionele voorbeeld van wit op swart as goeie kleurpaar gee hulle

ook nie verdere voorbeelde van spesifiek goeie kontrasterende kleure sodat dit duidelik is wat dit prakties beteken nie. Dié enkele voorbeeld is onvoldoende, want soos Kilgarriff en Kosem (2012: 49) aandui, is wit en swart lank reeds nie meer die enigste kleuropsies in moderne woordeboeke nie. Dit kan dalk selfs onbedoeld lei tot die voorkoms van 'n aanlyn woordeboek wat niks meer as 'n gedrukte ontwerp op 'n skerm is nie, waarteen Almind (2005: 37) waarsku. Die vroeër genoemde kwessie van estetika en die skynbare en inherente bruikbaarheid van 'n koppelvlak is ook weer ter sake. In die plek van hierdie tradisionele kombinasie verwys ander na rooi en groen as alternatiewe kleurpaar wat dikwels in gebruikerskoppelvlakke voorkom (vgl. Kilgarriff en Kosem 2012: 49; Jamil en Denes 2024: 53). Hiermee is daar egter ook 'n probleem, want rooi en groen is juis een van die algemeenste kleurpare wat vir veral kleurblinde gebruikers probleme kan veroorsaak (vgl. Soegoto, Fadhlurrahman en Hermawan 2019: 2).

Daar kan egter voorgestel word dat die kontras wat in aanlyn woordeboeke se kleurgebruik gehandhaaf moet word, op twee vlakke kan geld: eerstens tussen die voorgrond (al die tekstuele dele tesame) en die agtergrond, en tweedens tussen die tekstuele dele op die voorgrond. Betreffende die voorgrond-agtergrond-kontras moet kleure uiteraard maklik onderskeibaar wees sodat die data nie met die agtergrond versmelt nie. Dit blyk dat donker teks op 'n ligte agtergrond vir die meeste mense meer leesbaar is as ligte teks op 'n donker agtergrond — sien in hierdie verband Dobres, Chahine en Reimer (2017: 68-69) se verwysing na verskeie studies waarin dit bevind is. Verder kom dit voor asof swart of donkerblou teks op 'n wit agtergrond makliker lees en meer esteties is as ander kleurpare (Shieh en Lai 2008: 345). Beide hierdie toepassings behoort vir alle woordeboekgebruikers toeganklik te wees, want ook diegene wat soms sukkel om tussen kleurskakeringe te onderskei, kan kontras opmerk (Bigman, in Berisso 2018: 96).

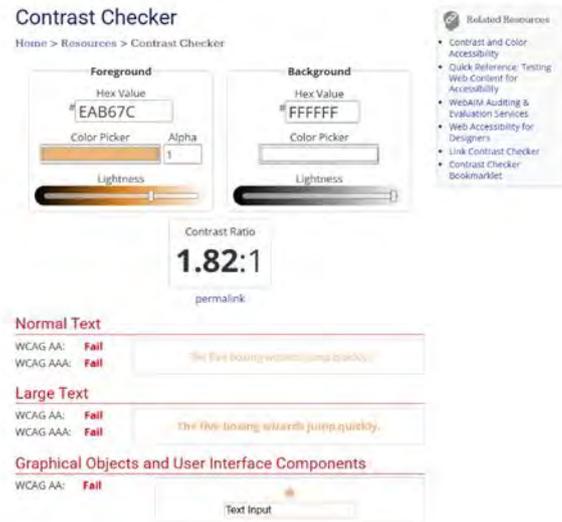
'n Oorweging van die kontras tussen die voorgrond en agtergrond alleenlik verseker egter nie dat alle woordeboekgebruikers die kleurgebruik funksioneel sal ag nie. Daar moet ook kontras tussen komponente van die gebruikerskoppelvlak op die voorgrond wees (Jamil en Denes 2024: 55), oftewel teks en komponente soos struktuurmerkers, aftrekkieslyste, hiperskakels en knoppies om data te versteek of te onthul. Die maklikste manier om te verseker dat alle woordeboekgebruikers tussen voorgrondkleure kan onderskei, is om sekere kleurkombinasies in die geheel te vermy (Berisso 2018: 96) en eerder 'n sogenaamde kleurblinde kleurpalet te gebruik. 'n Kleurblinde kleurpalet maak tegelyk vir alle woordeboekgebruikers voorsiening, dus sowel gebruikers wat sukkel om tussen bepaalde kleure te onderskei as diegene wat nie probleme daarmee ervaar nie (vgl. Soegoto et al. 2019: 5). Daar is dus geen party wat deur die toepassing daarvan benadeel word nie. Prakties behels dit die vermyding van rooi-groenkombinasies, of geel en blou saam omdat dit algemeen bekend is dat veral kleurblinde mense kan sukkel om tussen hierdie kleurkombinasies te onderskei (Berisso 2018: 94). Verder word helder en donker kleure in 'n kleurblinde palet in kombinasie gebruik sodat mense makliker tussen verskillende kleure kan differensieer (Soegoto et al. 2019: 5); die verskil tussen dié kleure is met die eerste

oogopslag meer opmerklik, en helder kleure is boonop makliker onderskeibaar as dowwer kleure, wat geneig is om te vermeng (Bigman, in Berisso 2018: 96).

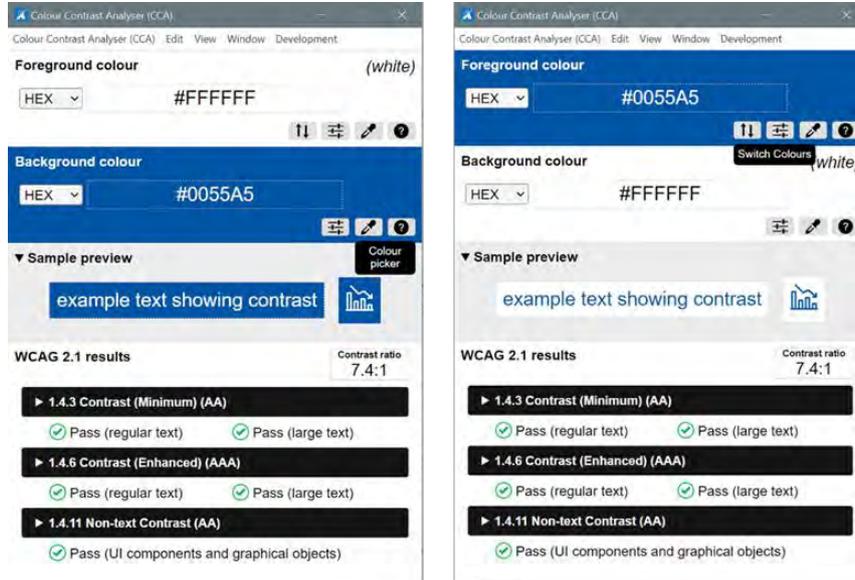
Met inagneming van die vroeër verwysing na aanlyn woordeboeke se aansluiting by webformateringsbeginsels, kan riglyne vir hoër inhoudstoeganklikheid op gebruikerskoppelvlakke geraadpleeg word om die gebruik van kontrasterende kleure meer konkreet te maak. Die Webinhoudtoeganklikheidsriglyne (Web Content Accessibility Guidelines in Engels, oftewel WCAG 2024) gee byvoorbeeld riglyne om toegankliker, meer gebruikersvriendelike webblaaie moontlik te maak en is gerig op kwessies wat met waarneembare en gebruikaspekte van webblaaie verband hou. Een tegniek vanuit die riglyne (sien riglyn 1.4.3 in WCAG 2024) is die gebruik van minstens 'n 4,5:1-kontrasverhouding tussen die teks en agtergrond by gewone teks (dit wil sê, teks kleiner as 18 pt indien dit nie vetgedruk is nie, of 14 pt indien vetgedruk), of 'n 3:1-verhouding by groter teks (groter as 18 pt, soos in opskrifte). Hierdie 3:1-verhouding behoort volgens die riglyne (1.4.11 in WCAG 2024) ook van toepassing te wees op komponente in die voorgrond, soos die teks en knoppies van die koppelvlak. Dit is slegs onnodig om hierdie voorgrondkontras te handhaaf wanneer komponente soos knoppies onaktief is (dus nie interaksie van die gebruikerskant af vereis is) of waar dit deur die gebruiker gewysig is om anders te lyk as wat die webontwerper in gedagte gehad het (WCAG 2024).

Dit is hier nodig om daarop te wys dat riglyn 1.4.3 en 1.4.11 tipies in samehang met 1.4.6 beskou word. Dié riglyn is 'n strengere toepassing van 1.4.3 en gerig op kontrastoeganklikheid vir gebruikers met geweldige sigprobleme. Toepassing daarvan sal vereis dat gewone teks 'n 7:1-kontras (in plaas van 'n 4,5:1-kontras, soos by riglyn 1.4.3) het, en groter teks oor 'n 4,5:1-kontras (teenoor die 3:1-kontras van 1.4.3) beskik. In woordeboeke waarin teks dikwels gekondenseerd en dig is, kan hierdie riglyn egter tans nog redelik ingewikkeld wees om te handhaaf. Daar word daarom hier voorgestel dat woordeboekopkoppelvlakontwerp aanvanklik op riglyn 1.4.3 en 1.4.11 gerig is, met latere toepassing van riglyn 1.4.6 indien nodig. Laasgenoemde word dus nie geïgnoreer nie, maar is dalk iets wat eerder mettertyd ingewerk kan word wanneer woordeboekplatforms werklik volwaardige aanlyn produkte word. Dit is dalk ook iets wat eerder met 'n skakelaar of skuifknoppie geaktiveer kan word sodat die gebruiker self kan bepaal of hulle teks en elemente volgens 1.4.11 geformateer wil hê.

Om hierdie riglyne se voorgestelde kontrasverhoudings prakties en maklik te kan meet, kan verskeie hulpbronne geraadpleeg word waarop 'n mens kleurkodes kan invoer om te kyk of dit aan die Webinhoudtoeganklikheidsriglyne sal voldoen. Dié platforms of toepassings gee dan vinnig terugvoer om te wys of die kleurpaar geskik kan wees of nie (sien Figuur 1 en 2 hieronder). Enkele voorbeelde sluit in TPGi se Colour Contrast Analyser (2025), WebAIM (2025), of Luminosity Colour Contrast Ratio Analyser deur Juicy Studio (2025), wat almal gratis is om te gebruik. Ander hulpbronne, ook gratis, gee weer breedvoerige voorbeelde en palette en stelle bypassende en webveilige kleure wat sigbaar behoort te wees — sien in hierdie verband byvoorbeeld die Index of Color Contrast Samples (s.a.).



Figuur 1: Die kleurpaar slaag nie die kontrastoets nie, soos bepaal deur WebAIM (2025) se aanlyn hulpbron



Figuur 2: Hier slaag die kleurpaar die kontrasriglyne, soos getoets op TPGi se Colour Contrast Analyser (2025) wat gebruikers toelaat om met 'n druppertjie die kleurpaar te toets, en maklik met die druk van 'n ander knoppie te ruil

Hierdie kwessie moet in verband gebring word met Müller-Spitzer en Kopleinig (2014: 164) se siening dat woordeboeke sover moontlik volgens 'n hindernisvrye ontwerp saamgestel moet word. Soveel moontlik gebruikers moet die koppelvlak se ontwerp dus toeganklik vind (Jamil en Denes 2024: 53), aangesien toeganklikheid direk tot funksionaliteit bydra. Dit is veral belangrik dat woordeboeke as gebruiksinstrumente vir alle gebruikers voorsiening moet maak indien moontlike aanpassings geen party benadeel nie.

3. Kleur in die HAT Aanlyn, Aanlyn WAT en WAV

In hierdie afdeling word enkele artikels uit drie aanlyn Afrikaanse woordeboeke — die HAT Aanlyn, Aanlyn WAT en WAV — ten opsigte van kleurgebruik bespreek, met enkele opmerkinge wat ook oor die strukture of uitleg gemaak word. Daar word primêr gefokus op hoe kleur reeds in die artikels gebruik word, maar waar gepas, word voorstelle vanuit die vorige afdeling betrek om dit duideliker te maak.

Die bespreking word nie gestruktureer ten opsigte van die genoemde woordeboeke nie, maar eerder volgens kategorieë: afkortings, uitdrukkings, en geëtiketteerde lemmas. Dit is so gekies sodat die kleurgebruik by verskillende datatipes oorweeg kan word. Aandag word nie aan die datatipes of inhoud geskenk nie, maar eerder hoe die datatipes aan die woordeboekgebruiker vertoon word.

3.1 Afkortings

Indien 'n afkorting soos QR in die WAV nageslaan word (Figuur 3), is die algemene indruk van die gebruik van kleur dat meestal duidelike kleure gebruik word. Verskillende kleure word aangewend om verskillende datatipes te merk. Dit sal vaardige woordeboekgebruikers toelaat om vinnig op die verlangde data te kan afkom indien hulle die woordeboek se kleurkodes leer ken.

QR afk. [Engelse verkorting van **quick response**]

› **QR-kode** | QR-kodes | soort tweedimensionele strepieskode wat bestaan uit swart blokkies van verskillende groottes op 'n wit agtergrond, soos afgedruk op papier of gegenereer op 'n skerm, van waar 'n skandeerder of slimfoon dit vinnig kan lees en dekodeer: › *By enige deur of hek tot enige gebou en by 'n hele paar ad hoc-beheerpunte wat oornag opspring rondom woonbuurte, moes almal weer 'n QR-kode (soms twee) skandeer met hul fone en hul temperatuur laat neem met 'n meetgewertjie* (Etienne Fourie, *Rapport*, 6 Februarie 2022). › *Skandeer die QR-kode vir 'n aanlyn toer van die eiendom* (Netwerk24, 4 Februarie 2022). › *Al hoe meer mense het oorgeskakel na digitale betaalmodes – hetsy tik-en-betaal, e-beursies of deur QR-kodes te skandeer* (Liesl Peyper, 19 Desember 2021).

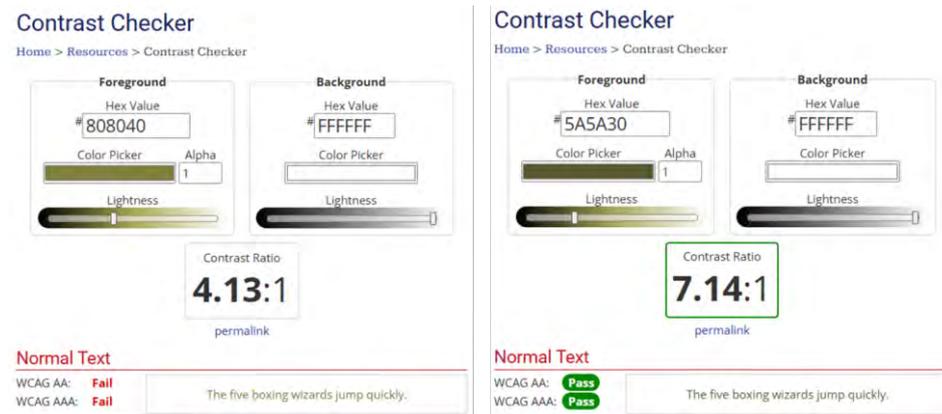
Figuur 3: QR soos in die WAV

Dit is opmerklik dat dieselfde donkerblou (en boonop vetgedrukte teks as pluspunt) gebruik word om die lemma as gidselement te merk, asook die volvorm van die leksikale item wat deur die lemma verteenwoordig word. Dit is nuttig en gerieflik, want woordeboekgebruikers wat slegs die volvorm wil naslaan, sal dit vinnig vind deur die oog na die bypassende blou in dieselfde reël te laat gly. Die lemma en volvorm word ook duidelik van die woordsoortaanduiding en aanduiding van die Engelse oorsprong geskei, wat elk in hulle eie kleure gemerk is. In die betekenisparafraseveld word 'n nuwe kleur nogmaals gebruik, wat die betekenisparafrase as nog 'n mikrostrukturele element duidelik van die omliggende datatipes differensieer. Ná die betekenisparafrase word drie sitate aangebied, met die sitaat en gepaardgaande bronverwysing wat in twee afsonderlike kleure gemerk word. Dit is baie nuttig, want die kleurkombinasie van die sitaat en bronverwysing is duidelik onderskeibaar en 'n gebruiker kan vinnig sien waar die sitate vandaan kom en hoe oud die sitaat is. Dit kan byvoorbeeld help om die gebruikskonteks te bepaal, of die geloofwaardigheid en betroubaarheid van die sitate te probeer beoordeel. Met inagneming van die sitaat se ouderdom kan die gebruiker weer iets oor die ontwikkeling of eietydse gebruiklikheid van die leksikale item aflei.

Deur telkens die verskillende datatipes in verskillende kleure te merk, word kleur as nietipografiese struktuurmerker ingespan. Dit beteken dat kleur as formele element in die artikel gebruik word, en 'n sterker identifiseringswaarde moontlik maak deur die soeksonde duideliker te merk. Dit versterk die punt van vroeër dat kleur tot 'n meer gebruikersvriendelike mikro-argitektuur kan bydra deur sekere elemente visueel van mekaar te onderskei en te blok. Soos genoem, kan kundige woordeboekgebruikers met 'n goeie woordeboekkultuur data só vinniger opspoor deur na spesifieke kleure te soek, met die voorwaarde dat 'n verduideliking van die betekenis van spesifieke kleure iewers vermeld moet word. Dit geld weer eens hier, want as kleur as struktuurmerker gebruik word, is dit nodig om dit byvoorbeeld in die gebruikersgids toe te lig wanneer die ander struktuurmerkers verklaar word.

Betreffende die sigbaarheid van die kleure is die meeste kleure donker, wat voordelig is met inagneming van die wit agtergrond. Daarbenewens is dit nuttig dat verskillende kleure, soos blou, grys, purper en 'n olyfgroen, gebruik word, asook kleure wat redelik maklik van mekaar onderskei kan word. By die gebruik van die ligter kleure kan daar egter dalk ietwat aangepas word. As die grys waarin die woordsoortaanduiding gemerk is met WebAIM (2025) teen 'n wit agtergrond getoets word, lewer dit 'n 6,79:1-kontras, wat bevredigend is met inagneming van die vroeër genoemde riglyne. Die olyfgroenkleur lewer weer 'n 4,13:1-kontras, wat nie aan die 4,5:1-mikpunt vir gewone teks voldoen nie. Dit beteken nie dat die kleur totaal onduidelik is nie, maar dit is eerder 'n aanduiding dat 'n effense aanpassing daarvan dit vir meer gebruikers duideliker sal maak. Sou die huidige olyfgroen byvoorbeeld effens donkerder gemaak word, soos met die skuifknoppie op WebAIM (2025), sal dit kan slaag sonder om die totale voorkoms drasties te moet verander (Figuur 4). Soortgelyk kan kleure op

die voorgrond, oftewel die verskillende tekstuele elemente se kleure, getoets word ten einde die kontraswaarde daartussen te bepaal en aan te pas, indien nodig.



Figuur 4: Die oorspronklike olyfgroen (links) kan maklik effens donkerder gemaak word om duideliker te wees (regs)

Wanneer die afkorting *URL* in die WAV nageslaan word (Figuur 5), is dit duidelik dat die kleurgebruik konsekwent toegepas word.

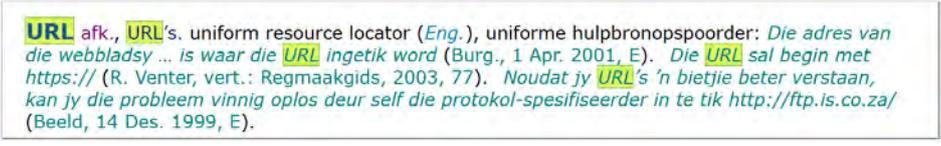
URL *afk., s.nw.* | URL'e, URL's | [Engelse verkorting van **Uniform Resource Locator**] (*internet*) adres wat 'n webblaaiër 'n spesifieke webblad of lêer op die internet help vind: › *As jy aanlyn koop, wees op die uitkyk vir verdagte URL's* (Huisgenoot, 20 November 2020). › *Die webleser gebruik die adres, vertaal die domeinnaam na die bediener se IP-adres en die res van die URL dui die pad na die spesifieke lêer op die gevindde bediener aan.*

Figuur 5: *URL* in die WAV

Die afkorting en volvorm is weer in dieselfde skakering blou, en ook dieselfde skakering as in die artikel van *QR*. Soortgelyk is die woordsoort, aanduiding van die afkorting se oorsprong, betekenisparafraze, en sate en gepaardgaande bronverwysings konsekwent gemerk. Dit is bevredigend, want soos genoem, sal dit die woordeboekgebruiker met 'n goeie kennis van die WAV help om makliker die verlangde inligting te onttrek. Wat egter effens onduidelik is, is waarom die etiket en aanduiding van die afkorting se Engelse oorsprong in dieselfde purper gemerk is wanneer alle ander datatipes afsonderlike kleure kry.

Wanneer dieselfde lemma in die Aanlyn WAT nageslaan word (Figuur 6),

is die voorkoms van die kleur met die eerste oogopslag ook redelik geslaag. Soos in die WAV, word die lemma, woordsoort, meervoudsvorm, volvorm en sitate byna elk in afsonderlike kleure gemerk.



URL afk., **URL**'s. uniform resource locator (Eng.), uniforme hulpbronopspoorder: *Die adres van die webbladsy ... is waar die URL ingetik word* (Burg., 1 Apr. 2001, E). *Die URL sal begin met https://* (R. Venter, vert.: Regmaakgids, 2003, 77). *Noudat jy URL's 'n bietjie beter verstaan, kan jy die probleem vinnig oplos deur self die protokol-spesifiseerder in te tik http://ftp.is.co.za/* (Beeld, 14 Des. 1999, E).

Figuur 6: Die Aanlyn WAT se artikel vir *URL*

Hier word die volvorm egter as betekenisparafrase beskou en in swart gegee, wat teenoor die blou, vetgedrukte lemma en meervoudsvorm staan. Relatief tot die WAV sal die benadering effens meer moeite van die woordeboekgebruiker vereis, want die gebruiker kan nie in dieselfde reël net aanhou soek na die blou waarin die lemma gegee is om die volvorm te kry nie. Die gebruiker moet eerder die besonderhede lees om dit te bekom. Nietemin is die woordsoort wat in pienkpers gemerk is duidelik onderskeibaar van die lemma en meervoudsvorm. In die betekenisparafrase is 'n ander voordeel dat die presiseringkruisverwysings wat deur hiperskakels gevolg kan word, gemerk is in die blou kenmerkend van hiperskakels. In dié wyse voldoen die Aanlyn WAT aan Müller-Spitzer et al. (2012: 441) se reeds genoemde punt dat woordeboekgebruikers wil hê aanlyn woordeboeke moet 'n bekendheid met ander toepassings en webformateringsbeginsels toon. Woordeboekgebruikers behoort die hiperskakels danksy die tipiese blou onmiddellik as sodanig te kan sien, wat gerieflik is en toeganklikheid verhoog.

Iets wat nie in die WAV gedoen is nie, maar hier 'n aanwinst is, is die groen blok om die leksikale item wat nageslaan is. Dit laat woordeboekgebruikers toe om vinnig deur die artikel te kan gaan en die gesoekte item te kan raaksien, ongeag waar dit in die artikel staan. In die geval van 'n naamwoord soos *URL* is dit nuttig as 'n gebruiker iets oor die sintaktiese optrede wil kan aflei deur byvoorbeeld te kyk watter lidwoord(e) voor die item kan staan, of in ander items se geval dalk watter setsels daarmee optree, of 'n adjektief attributief of predikatief optree, ensovoorts. Daar is hier egter twee beperkinge. Die woordeboekgebruiker het eerstens nie die opsie om die blok te verwyder nie. Dit is egter wenslik om byvoorbeeld 'n knoppie te kan druk wat die blok afhaal of die opsie geheel afsit sodat 'n skoner weergawe van die artikel telkens vertoon word. Die tweede beperking is dat die groen blok slegs om die gesoekte vorm van die lemma verskyn. In die geval van byvoorbeeld 'n werkwoord sal slegs die teenwoordige tyd gemerk word as die woordeboekgebruiker dit so gesoek het; enige ander vorm van dieselfde item (soos die verledetydsvorm of die leksikale item as deelwoord) bly ongemerk. As die gebruiker spesifiek net daardie vorm soek, is dit nuttig. Dit sou wel nut-

tiger wees om moontlik 'n tweede soekmoontlikheid in te sluit wat alle vorme van die item merk om ook daardie vorme wat nie ingetik is nie, in te sluit.

Wat die WAT-redaksie verder kan oorweeg, is om soos in die WAV die bronverwysings in ander kleure as die sitate te merk. Die WAT is 'n omvattende woordeboek, en omvattendheid behoort ook op mikrostrukturele vlak en binne elke mikrostrukturele element te geld (Gouws 2017: 442). Daarom kan dit nuttig wees om bronverwysings in ander kleure te merk omdat dit as visuele skeiding tussen sitate kan dien en gebruikers sal toelaat om vinnig die ouderdom en bron van elke sitaat vas te stel. Hoewel die WAT 'n sinchroniese woordeboek is, beskik dit as langtermynprojek ook in 'n mate oor diachroniese kenmerke. Dit is tipies van omvattende woordeboeke dat daar 'n mate van 'n chronologiese aanduiding van vorm- en betekenisontwikkeling is (Gouws en Prinsloo 2005: 49). Deur dan bronverwysings duideliker van die sitate te onderskei, kan woordeboekgebruikers met 'n oorsig die diachroniese gegewens raaksien, wat behulpsaam kan wees.

3.2 Uitdrukings

Die artikels wat hierbo gewys is, is telkens kort. In sodanige artikels is dit waarskynlik vir woordeboekgebruikers makliker om die verlangde data te vind, ongeag die kleurgebruik, omdat daar gewoon minder data is wat hulle moet deurwerk. In langer artikels wat veral algemeen is in 'n omvattende woordeboek soos die Aanlyn WAT kan dit egter ingewikkelder wees, en 'n gebruikersvriendelike mikroargitektuur (wat sinvolle kleurgebruik insluit) is dan juis noodsaaklik. Dit is omdat langer artikels uiteraard meer data bevat en 'n hoë kwantiteit data wat swak bestuur word, kan daartoe lei dat woordeboekgebruikers data-oormoed ervaar (Gouws en Tarp 2017: 392). Dit vind plaas wanneer iemand se eiesoortige kapasiteit om inligting te verwerk en in te neem, oorskry word (Speier, Valacich en Vessey, in Gouws en Tarp 2017: 395). Die gevolge is dat gebruikers angs of verwarring ervaar, en hulle onttrek dan nie die benodigde inligting nie omdat dit waarvan hulle probeer sin maak ongeorganiseerd, onbeheersd en inkoherent voorkom (Wurman en Bradford, in Gouws en Tarp 2017: 395). Die artikels kom dan opgeblaas voor en raak weens 'n oormoed data onprakties (Gouws en Tarp 2017: 392). Dit verhoed direk dat 'n woordeboek aan sy doel as praktiese gebruiksinstrument kan voldoen.

Indien *soveel* byvoorbeeld in die Aanlyn WAT nageslaan word (Figuur 7), behoort dit uit die uitdrukingsrubriek duidelik te wees waarom die kwessie van data-oormoed belangrik is om te oorweeg in teksdigte, gekondenseerde artikels. Hier is dieselfde lof oor duidelike kleurgebruik wat vroeër genoem is nie sonder meer van toepassing nie. Die uitdrukingshofie (asook die twee subhofies "gespesialiseerd" en "idiomaties") is duidelik in swart gemerk en kan maklik gedifferensieer word van die kernkomponent as soekwoord, maar van daar af sal die data onder die subhofies nie vir alle woordeboekgebruikers ewe duidelik onderskeibaar wees nie.

UITDR.
A GESPESTALISEERD
moontlik: soveel as moontlik Ook, geselstaal, *suffel* as *moontlik*. Die maksimum getal of hoeveelheid, of die maksimum waartoe jy in staat is of wat uitvoerbaar is: 'n Kind oriënteer hom ... in die wêreld deur vertrou te raak met **soveel** as moontlik van die voorwerpe ... wat hy ... teekom (H.J. Harmse in Harmse – Kirstein: *Redag.*, 2, 1980, 233). *Eenkeer, toe hy my in sy tuin betrap waar ek in my vuil néppie sit en tuinslakke kou, het hy my aangerai om soveel as moontlik slakke te eet, want ... die goed is 'n pes*" (K. Kombuis in Insig, Des, 2001, 61). *Die Hoofbestuur het ... aan sy verteenwoordigers in die Provinsiale Adviserende Komitee Insake Onderwysdiens (PADKO) opdrag gegee om te beding om soveel as moontlik onderwysposte te behou* (Unie, Sept. 1991, 54).

soveel: soveel as (n.d. Eng. as much as)
1 Ook, geselstaal, *suffel* as. Die maksimum of tot 'n maksimum van: **Soveel** as wat 'n man op 'n leerwa kon laai, is vir een pond (R2) verkoop (J.D. Burger in F.A. Venter: *Water*, s.j., 430). *Dit (die elektroniese mikroskoop) kan voorwerpe net 'n paar maal of soveel as 1,600,000 maal vergroot* (Burg., 30 Jun. 1998, 5 --- opskr.).
2 Ook, geselstaal, *suffel* as. Eens of enigins: Daar is nie soveel as 'n groen uielof vir slaai nie (E. Kotze: *Wind*, 2007, 73). *Hulle kon hom nie 'n tree versit nie --- hulle kon hom nie soveel as roer nie* (C.J. Langenhoven: *Werke II*, 1956, 25).

3 Ook, geselstaal, *suffel* as. Niet (*NET 1) of bloot (*BLOOT 5): *As sy Lordskap soveel as 'n ooglid sien beef solank as hy hierdie skepsel galg toe dons, sal hy julle laat slaan met die kats* (C.J. Langenhoven: *Weg* [2], 1950, 103). *Die Liewenheer het my geseën met die hardste baard denkbaar, ... want as ek soveel as een enkele dag oorskraan, lê die skermestom se snykant eers om en klim dan kou-kou bo-op die stoppels op sonder om een af te kry* (P.H. Nortje: *Spook*, 1995, 7).

4 Ook, geselstaal, *suffel* as. In dieselfde mate as: *Erna is deel van myself. Soveel as wat ek haar handewerk is, soveel is sy myne* (A.H. de Vries: *Kruispad*, 1966, 9). *Intussen moes die reis voortgaan, en moes sy voorgee om iets wat alle pit verloor het, nog net soveel as voorheen te geniet* (S. Troskie: *Rapsodie*, 1962, 186).

5 Ook, geselstaal, *suffel* as. So baie as: "Alto, oom Gert --- soveel as ek van 'n advokaat dink --- hoop ek tog van harte dat ek hom nooit in my lewe nog nodig kry nie" (E. Elff-v.d. Walt: *Fontein*, 1956, 77). **Soveel** as wat sy van babas hou, kan sy nie eers begin om haar voor te stel hoe dit moet wees om twee pasgeborenes gelyk te versorg nie (E. Steyn, vert.: *Skeuring*, 2011, 1).

soveel: soveel so (n.d. Eng. so much (so) that) Ook, geselstaal, *suffel* so. In so 'n mate: *Van die kerkvaste Afrikaners in Rhodesië is dié geslag wat in die land gebore is --- soveel so dat hulle dikwels nie hul kerk verlaat wanneer hulle met Engelsprekendes in die huwelik tree nie* (Huisg., 21 Jan. 1966, 14). *Fiat ... se aandeel in die Europese notamark het sedert 1984 standhoudend gegroei. Soveel so dat hy Volkswagen verlate jaar as die markleier in daardie mark oortref het* (Tegniek, 21 Jul. 1989, bylae, 12). Vgl. SOSEER 1.

soveel: soveel te ... Ook, geselstaal, *suffel* te ... In 'n hoë mate: *Almal (wat wil gaan boer) moet in die geleentheid wees om eers so ver te kom in die skool as moontlik. As almal eers kan matrikuleer, soveel te beter* (M.-L. Theron: *Reenboog*, 1940, 114). *Hy wil vir eens en vir altyd ontslae raak van die ander man se telg in sy huis. En as dié telg nog 'n plaas en 'n flukse skoorsuin by Hagelkraal se bates kan voeg, soveel te beter* (S.D. Fourie: *Lantemswaier*, 2011, 149). *Die oprigting van 'n kerk was soveel te meer noodsaaklik vir diegene wat op die afgeleë grens gewoon het* (C.G. Botha: *Ons S.A.*, s.j., 32).

soveel: soveel vir (n.d. Eng. so much for) Ook, geselstaal, *suffel* vir. Tot daarnatoe (²DAARNATOE) met: 'n Groot dankie aan mnr. J.P. vir sy daadwerklke bydrae om die Sionskerk te restoreer. Dit sou hom seker heelwat minder gekos het as hierdie gebou ... nie deur vandale bygekom is nie --- soveel vir die bewaringsgedagte wat sommige Paarlette het (Burg., 23 Mrt. 1983, 22). Toe die verkiesing verby is, skop hulle die AWB in die tande en later selfs vir die KP. **Soveel** vir die samewerking aan die regterkant (K. Smith: *AWB*, 1996, 87).

B IDIOMATIES
soveel: soveel en soveel (soms *suffel* en *suffel*) Met 'n eienskap, hoeveelheid of graad wat nie gespesifiseer word nie: *Dis 'n vloedjaar --- amper weektik word berig van mense wat dakloos gelaat is, oor soveel en soveel miljoen rand se skade* (T. Holland: *Nog 'n casanova*, 2016, 1).

soveel: soveel (soms suffel) te sê (n.d. Eng. so much as to say) Asof om te sê: *As ons dít (die parlement) die praathuis noem, is dit gewoonlik met 'n spotlaggie, soveel te sê: die lede verkies hul tyd met praat in plaas van hul werk te doen en wette te maak* (Herman – Goldblatt: *Janus*, 1962, 54).

Figuur 7: Die Aanlyn WAT se hantering van uitdrukkings onder *soveel*

Soos hierbo te sien, word drie kleure gebruik om onderskeidelik die volgende te merk: 'n mosgroen kleur vir die uitdrukking en betekenisparafraze, 'n asuurblou vir die wisselvorm, en 'n blougroen vir die sitate. Relatief tot mekaar lyk die kontrasverhoudings soos volg indien die kleurcodes op WebAIM (2025) ingevoer word: 1,49:1 vir die mosgroen teenoor asuurblou; 1,36:1 wanneer die mosgroen teenoor die blougroen gemeet word; en 1,09:1 vir die asuurblou relatief tot die blougroen. Hierdie groenblou palet kan esteties wees deurdat dit pas by die WAT se kenmerkende blou palet, waarna vroeër ook verwys is, maar dit misluk redelik in die 4,5:1-mikpunt vir gewone teks. Dié mikpunt is wel net 'n riglyn, maar die lae kontras tussen die gebruikte kleure dui daarop dat minstens sommige gebruikers, veral kleurblinde of siggestremde gebruikers, dit uitdagend kan vind om gemaklik en met sekerheid tussen hierdie kleure te onderskei. Dit lei weer tot 'n minder effektiewe en praktiese gebruik van kleur wat veroorsaak dat dit 'n woordeboekgebruiker langer kan neem om deur die data te werk omdat hulle nie met dieselfde gemak as voorheen net direk die tersaaklike deel kan identifiseer deur 'n spesifieke kleur te soek nie.

Dit kan uiteindelik in artikels so lank soos dié van *soveel* (en veral in langer artikels omdat die Aanlyn WAT ook baie daarvan bevat) moontlik tot data-oormoedigheid lei. Dit is omdat baie data in die artikel gehuisves moet word, maar nie noodwendig goed bestuur word in terme van hoe dit uitgelê is en van omliggende data geskei word nie. Die vroeër punt oor 'n dinamiese veelvlakkige artikelstruktuur geld weer hier, want in lang artikels soos in die Aanlyn WAT kan 'n soektonnel en die gebruik van beperkte artikels help om data-oormoedigheid te verminder. Daarbenewens kan dit woordeboekgebruikers self laat besluit in watter mate hulle verrykte en uitgebreide data wil bekom. Soos genoem, kan kleur so ingespan word om woordeboekgebruikers te wys waar hulle die behoefte op

meer data moet aandui om sodoende deur die soektonnel te beweeg. Dit kan ook gebruik word om data binne die soektonnel met ander dele van die woordeboek in verband te bring.

Indien die wisselvorm *suffel* nageslaan word (Figuur 8), is die situasie ietwat beter. Hier is die risiko van data-oorlading effens minder en dit behoort makliker te wees om tussen datatipes te onderskei.

suffel (suf-fel) (geselstaal)

I telw.

1 Sien SOVEEL I 1: *Suffel* reën soos nou het ek in '61 laas gesien (hulsg., 19 Des. 1985, 111). Dit is *suffel* (beduie met duim en voorvinger) swakker as laasweek (Carl Boplaas: Knikkertjie, 1998, 39). Vir my was dit 'n allerverskriklike bedrag. Waar sou ek *suffel* geld vandaan kry? (V. Schaikwyk – Hienstra: Duineboerdery, 2013, 249).

2 Sien SOVEEL I 2: Ek het jou nie *suffel* kere kraai toe sien loop nie (A.A.J. v. Niekerk: Bittergousblom, 1974, 8). Almal hier in Frik du Preezstraat werk mos alles uit volgens 1994. Dis altyd *suffel* jaar voor 1994 of *suffel* jaar daarna (J. Goosen: Pappaw, 2002, 12). Die Hollands wat hulle daai se dae gepraat het, was al van *suffel* skroeiwarm somers gelouter, dat dit toe al lekker winddroog was (J. le Roux: Engel, 2014, 186).

II bw.

1 Sien SOVEEL II 1: Na 'n ruk voel ek *suffel* beter lat ek ontstaan en met my twak afsit stal toe (Carl Boplaas: Knikkertjie, 1998, 23). O'Neill het hulle ... aangerysde brood bekendgestel as 'n voorbeeld van hoe geloof soos die onsigbare suurdeeg *suffel* meer van die koringkorrel maak (J. le Roux: Engel, 2014, E).

2 Sien SOVEEL II 2: Hy het in die distrik geboer. *Suffel* kan ek onthou. is "Hy's tog seker nie verlies op haar nie, dus hoef die gedagte jou nie te ontstel nie." "My dier, ek kan nie *suffel* omgee nie" (J. Meintjes: Mallemeule, 1964, 60).

UTDR.

A GESPESTALISEERD

moontlik: suffel as moontlik (geselstaal) Sien SOVEEL AS MOONTLIK (SOVEEL II, uitdr.): 'n Mens moet *suffel* as moontlik spaar vir jou oudag.

suffel: suffel as (geselstaal) (n.d. Eng. as much as)

1 Sien SOVEEL AS 1 (SOVEEL II, uitdr.): Wildeels-poeler. *Suffel* as wat op 'n halfkroon kan gaan, met water (D. Richard: Ysterhout, 1974, 5).

2 Sien SOVEEL AS 2 (SOVEEL II, uitdr.): Daar mag nie *suffel* as 'n stoffie of 'n strepie op wees nie (J. Spies: Profeet, 1984, 15).

3 Sien SOVEEL AS 3 (SOVEEL II, uitdr.): As jy *suffel* as jou oog knip, stuur ek jou uit die klaskamer uit.

4 Sien SOVEEL AS 4 (SOVEEL II, uitdr.): Jy skinder nog net *suffel* as altyd.

5 Sien SOVEEL AS 5 (SOVEEL II, uitdr.): *Suffel* as wat ek hom jammer kry, kan ek hom nie nou help nie.

suffel: suffel so (n.d. Eng. so much (so) that) (geselstaal) Sien SOVEEL SO (SOVEEL II, uitdr.): My kop het so gepyn --- *suffel* so dat ek skoon dubbelvisie gehad het.

suffel: suffel te ... (geselstaal) Sien SOVEEL TE ... (SOVEEL II, uitdr.): Ek is alleen; *suffel* te meer nou dat dit grendeltyd is.

suffel: suffel vir (n.d. Eng. so much for) (geselstaal) Sien SOVEEL VIR (SOVEEL II, uitdr.): My buurman leen nie sy grassnyer uit nie. *Suffel* vir buurmanskapt

B IDIOMATIES

suffel: suffel en suffel (meer dikw. soveel en soveel) (geselstaal) Met 'n elenskap, hoeveelheid of graad wat nie gespesifiseer word nie: Ag oom, ek se ook altyd ... die rokkie kos *suffel* en *suffel* maar dis nie waar nie (Mikro: Koperkanon, s.j., 78). Daai galjoentjie is net *suffel* en *suffel* werd (A.H. de Vries: Dubbeldoor, 1963, 79). Onopvallend op die tweede laaste bladsy 'n berig oor die aandklokkeëling ingevolge artikel *suffel* en *suffel* (J. Goosen: Kat, 1986, 58).

sé: suffel (meer dikw. soveel) te sé (geselstaal) (n.d. Eng. so much as to say) Asof om te sé: *Suffel* te sé: Hel, ek begeer dit nou diep in my hart ... (Carl Boplaas: Knikkertjie, 1998, 199). My ma en ant Mietjie kyk hom kwaad aan, *suffel* te sé: "Sé net iets, Koois Brits!" (J. Goosen: Ons, 1990, 142).

Figuur 8: Die Aanlyn WAT se hantering van *suffel*

Hier word die etiket "geselstaal" telkens in purper aangedui, wat dit goed onderskei van die omliggende teks en wat ook heelwat sigbaarder is. Om seker te maak dat dit sluit, word die etiket in *soveel* se artikel ingesluit, maar daar word dit in die betekenisparafraze as parentese gedoen. In terme van kleur word dit weer in mosgroen gedoen, wat dit minder sigbaar maak. Laasgenoemde maak wel sin met oorweging van die Aanlyn WAT se benadering, maar die gebruik van 'n ander, maklik onderskeibare kleur om 'n etiket in die minder gebruiklike wisselvorm "suffel" se artikel aan te dui, is baie nuttig.

Die verhouding tussen die artikel vir *soveel* en *suffel* word verder in die uitdrukkingsrubriek gesluit deurdat daar na "soveel" gepresiseer word, wat gepas is wanneer 'n woordeboekgebruiker na die meer gebruiklike wisselvorm verwys moet word. Wat dit egter vir sommige gebruikers ingewikkeld kan maak, is dat hierdie presisering in 'n redelik soortgelyke blou gedoen word as die hiperskakelverwysings bo die uitdrukkingsrubriek. Gebruikers wat sukkel om tussen sekere kleure te onderskei, kan dan dalk verward wees wanneer hulle op die teks, byvoorbeeld "SOVEEL AS 1 (SOVEEL II, uitdr.)", probeer klik en niks gebeur nie. Die kenmerkende blou waarin hiperskakels gemerk word, is dus goed, maar wanneer daar in 'n soortgelyke kleur gepresiseer word sonder 'n hiperskakel, kan dit minder effektief wees. Die soortgelyke kleure verg op die een plek 'n interaksie van die gebruiker en laat op 'n ander plek nie so iets toe nie.

As bogenoemde kleurkombinasies aangepas moet word om vir meer gebrui-

kers duidelik te kan wees, kan dit soos by die WAV se olyfgroen van vroeër ietwat aangepas word deur die kleure se helderheid (hoe lig of donker dit is) te verander, of 'n bietjie met die saturasie te speel. Andersins, en dit is hier waarskynlik die makliker roete, kan daar herbesin word oor die Aanlyn WAT se kleurgebruik en met 'n kleurblinde palet in gedagte beplan word. Dit beteken dat die WAT se bloue grotendeels behou kan word en goed saam met oranje, rooi of bruin gebruik kan word, maar nie so goed saam met pers, groen of dowwe grys nie (vgl. Soegoto et al. 2019: 4-5). Dit is boonop reeds duidelik in hoe goed die kleur van die etiket werk, asook hoe die kombinasie van die valer groene en bloue minder geslaagd is.

Wanneer die uitdrukkingsveld in die HAT Aanlyn oorweeg word, byvoorbeeld in die artikel van *vlieg* (Figuur 9), is dit dadelik opmerklik dat 'n eenvoudiger, skoner ontwerp gevolg is.

vlieg? ww. [het **gevlieg**]

1 deur middel van vlerke in die lug opstyg en voortbeweeg: *voëls/insekte/vlermuise vlieg*.
2 (*lugvaartkundig*) met 'n vliegtuig reis, daardeur vervoer word of as loods dit self stuur: *Ons het direk van Kaapstad na Amsterdam gevlieg*. • *Die slagoffers van die busongeluk is per helikopter hospitaal toe gevlieg*. • *Noodvoorrade is na die dorpies gevlieg wat deur die oorstromings afgesonder is*. • *Ek wil leer vlieg*.

3 (*figuurlik*) vinnig beweeg/verbygaan/verdwyn: *Die skelm vlieg om die hoek*. • *Die skrywer se eerste boek het van die boekwinkelrakke gevlieg*. • *Die ruit het aan skerwe gespat toe die krieketbal daardeur vlieg*. • *As 'n mens hard aan die werk is, vlieg die ure behoortlik*. ◊ *Die dae van ons jare ... gou gaan dit verby en ons vlieg daarheen* (Ps. 90:10).
| *die tyd vlieg* die tyd gaan gou verby.

Uitdrukkings
mekaar in die **hare vlieg**
rusie maak; stry.
hoër vlieg as wat jou vlerke lank is
meer onderneem as wat jy kan behartig.
te hoog vlieg
bo jou vermoë lewe.
in die lug vlieg
ontplof.
► **vlieg**: **vliegduur**, **vliegeskader**, **vlieggebied**, **vlieghoogte**, **vliegklub**, **vliegles**, **vliegroete**, **vliegsnelheid**, **vliegsport**, **vliegtog**, **vliegtyd**, **vlieguitrusting**, **vliegvrees**

Figuur 9: 'n Gedeelte van *vlieg* se artikel in die HAT Aanlyn

Hierdie eenvoudiger, skoner ontwerp is uiteraard in 'n mate omdat die HAT Aanlyn as handwoordeboek nie so omvattend hoef en behoort te wees soos die Aanlyn WAT nie. Sonder die insluiting van sitate by uitdrukkings is dit ook makliker om 'n gebruikersvriendelike mikro-argitektuur te bewerkstellig. Nietemin is daar goeie beginsels toegepas wat die artikel toeganklik maak. Die groen blok om die gesoekte woord is reeds bespreek soos dit in die Aanlyn WAT voorkom, maar dit is weer hier noemenswaardig en nuttig. Dit het egter soos by die Aanlyn WAT dieselfde beperkinge. Wat die HAT Aanlyn se eiesoortigheid betref, is dat die purper kleur waarin die uitdrukkings gegee word heel onderskeibaar is van die swart waarin die uitdrukkinghofie en betekenisparafrase van elke uitdrukking aangebied is.

Iets wat hierdie benadering nog beter maak, is dat die purper redelik verskil van die ander kleure in die artikel, met uitsondering van die rooierige kleur van die lemma en nommers wat die betekenisonderskeidinge merk. Laasgenoemde is egter nie 'n probleem nie, want die purper staan nie direk langs of naby genoeg aan die rooi vir die kleure om verwarrend te wees nie. Betreffende die ander kleure is dit opmerklik dat swart, blou, oranje en olyfgroen gebruik word, wat almal kontrasgewys redelik van purper verskil. Gevolglik sou daar steeds maklik tussen die data onderskei kon word selfs al was meer data, soos kollokasies en/of sitate, ingesluit. Daar kan aangeneem word dat kollokasies en/of sitate in blou gegee sou word ter aansluiting by hoe dit in die res van die artikel gedoen is. Indien wel, sou dit steeds duidelik onderskeibaar wees van die uitdrukking in purper en die gepaardgaande betekenisparafrase in swart. Die probleem in die Aanlyn WAT se hantering van uitdrukkings is dus nie uitsluitlik vanweë die kwantiteit data nie, maar in 'n mate weens 'n wanbestuur, of eerder swakker aanbieding, van die data.

3.3 Geëtiketteerde lemmas

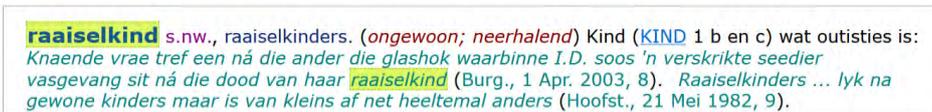
In al drie woordeboeke kan die gebruik van kleur 'n waardevolle rol by die oordrag van pragmatiese inligting speel deurdat dit etikette duideliker sigbaar kan maak. Etikette speel 'n belangrike rol omdat dit enige afwyking van die standaardtaal aandui, byvoorbeeld ten opsigte van formaliteit, gebruiksfrekwensie of gebiedsverankering, of andersins nuttige en nodige gegewens oor die gebruiksfere oordra. Hierdie gegewens kan by sekere leksikale items veral belangrik wees, en 'n woordeboekgebruiker wat nie duidelik genoeg van die etiket bewus gemaak word nie, kan byvoorbeeld in die verleentheid gestel word wanneer hulle 'n item as neutraal interpreteer wanneer dit eintlik minder gepas is vir 'n bepaalde situasie of selfs kwetsend is. Kleur kan juis hier gebruik word om dié beperkinge op die gebruik van 'n leksikale item verteenwoordig deur die lemma vir woordeboekgebruikers duideliker aan te dui.

Dit is byvoorbeeld duidelik hoe etikette in die WAV gemerk word, soos in die artikel vir *moffie* (Figuur 10). In dié artikel word *moffie* duidelik as informeel geëtiketteer in die bepaalde betekenisonderskeidinge wat in die WAV opgeneem word. Die gebruik van purper laat die etiket beter uitstaan relatief tot sowel die blou van die lemma en meervoudsvorm wat reg voor dit verskyn as die swart waarin die betekenisparafrase kort daarna gegee word. Daarbenewens laat die kleur die etiket duidelik by die ander kontekstuele leiding aansluit, want so ook is dieselfde purper gebruik om "deur lede van die LGBTQIA-gemeenskap gebruik om na hulle self te verwys" en "dikwels kwetsend, neerhalend deur cisgender mense gebruik in verwysings na lede van die LGBTQIA-gemeenskap" te merk. Hierdie besonderhede is belangrik om uit te lig en dui duidelik die gebruiksbepalinge aan om te wys dat die leksikale item "onder" die standaardtaal is en nie as neutraal geag moet word nie.



Figuur 10: Die artikel vir *moffie* soos in die WAV

Soortgelyk word etikette goed met kleur in die Aanlyn WAT uitgelig, soos genoem. Dit is ook hier goed in gevalle waar 'n lemma 'n potensiaal neerhalende leksikale item verteenwoordig, soos in die Aanlyn WAT se hantering van *raaiselkind* (Figuur 11).



Figuur 11: Die Aanlyn WAT se benadering tot *raaiselkind*

In só 'n geval kan woordeboekgebruikers van 'n omvattende woordeboek egter wonder wat in die Aanlyn WAT as neerhalend beskou word, of verdere toeligting oor etikette en hulle gebruik verlang. Hulle kan moontlik ook wonder waarom sekere items, veral rassismes, aansienlik korter en minder omvattend as ander items bewerk word. Vir hierdie gebruikers kan dit nuttig wees as kleur dan gebruik word om die verhouding tussen die etiket en die bepaalde buiteteks, in die Aanlyn WAT se geval die hanteringsbeleid vir beledigende en sensitiewe leksikale items, duidelik te maak. Dit kan op 'n paar maniere gedoen word, maar die etiket kan byvoorbeeld in een kleur gemerk word wat die gebruiker via hiper-skakelroete na die bypassende buiteteks neem wat met dieselfde kleur omraam word om die relevante teksdeel uit te lig. Nog 'n moontlikheid is 'n opskietspyskaart wat aan die kant van die skerm vertoon word, met oortjies na verskillende buitetekste in die WAT. Die oortjie na die hanteringsbeleid as tersaaklike buiteteks kan dan byvoorbeeld in dieselfde kleur gemerk word. Andersins kan 'n knoppie of ander simbool in een kleur binne die artikel gemerk word, met so 'n tipe oortjie in dieselfde kleur. Sodoende kan woordeboekgebruikers bewus gemaak word van waar hulle addisionele toeligting kan bekom om die woordeboek se spesifieke benadering beter te verstaan.

4. Gevolgtrekking

Uit die bespreking van die enkele punte is dit duidelik dat kleur 'n belangrike rol in aanlyn woordeboeke se artikel kan en moet speel. Dit is omdat dit toegang kan verbeter en 'n wesenlike bydrae tot 'n gebruikersvriendelike mikro-argitektuur kan lewer deur as nietipografiese struktuurmerker ingespan te word. Kleur alleenlik kan nie hierdie werk verrig nie, en moet ook nie die enigste middel wees waardeur bepaalde gegewens oorgedra word nie. Desondanks is dit iets wat 'n rol kan speel in hoe suksesvol 'n woordeboek as praktiese gebruiksinstrument kan dien. Daarom is dit nodig om doelbewus te dink oor die gebruik van kleur en die moontlikhede wat dit in die artikelstruktuur inhou, asook oor hoe die gebruik duidelik iewers toegelig moet word sodat woordeboekgebruikers dit kan verstaan. Heelwat van die moontlikhede in die teoretiese oorsig is egter tans nog toekomsmusik in die Afrikaanse leksikografie, soos die veelvlakkige artikelstruktuur. Dit is nietemin opmerklik dat kleur in 'n woordeboek soos die WAV waaraan werk meer onlangs begin is reeds doeltreffend begin gebruik word, selfs al is daar nog by oorweging van die individuele kleure 'n bietjie ruimte vir verbetering. Die ontwikkeling is wel belowend en hou opwindende moontlikhede in.

Indien die Webinhoudtoeganklikheidsriglyne oorweeg word, is dit duidelik dat kleur 'n rol in sensitiwiteit kan speel en redaksies kan help om hulle begrip van sensitiwiteit te verbreed. Wat hiermee bedoel word, is dat sensitiwiteit nie net ten opsigte van pragmatiese inligting oor bepaalde leksikale items hoef te geld nie. Dit kan ook geïnterpreteer word ten opsigte van bepaalde gebruikersbehoefte en -vermoëns. Dit is veral moontlik en nodig in aanlyn woordeboeke waarin kleiner aanpassings geen party sal benadeel nie, maar wel vir dié gebruikers wat dit nodig het 'n verskil kan maak sodat 'n woordeboek vir daardie gebruikers as praktiese gebruiksinstrument kan dien.

Eindnote

1. Soos aangekondig in 'n nuusbrief van 9 April 2025, het die Aanlyn WAT 'n nuwe baadjie gekry, wat onder meer 'n nuwe kleurskema insluit. Die skermskote wat in hierdie artikel ingesluit is, is egter reeds in Februarie 2025 geneem. Die nuusbrief is hier beskikbaar:
<https://www.wat.co.za/wat-wys-sy-kleure/>

Bronnelys

Woordeboeke

Aanlyn WAT: *Woordeboek van die Afrikaanse Taal.*

<https://www-woordeboek-co-za/> [20 Februarie 2025]

HAT Aanlyn: *Handwoordeboek van die Afrikaanse Taal.*

<https://mml-hat-co-za/> [24 Februarie 2025]

WAV: Woordeboek van Afrikaans Vandag.

<https://afrikaans-vandag.co.za/> [20 Februarie 2025]

Ander bronne

- Almind, R.** 2005. Designing Internet Dictionaries. *Hermes, Journal of Linguistics* 18(34): 37-54.
- Berisso, K.** 2018. Addressing Color Blind Awareness in the Classroom. *Journal of Business and Management Sciences* 6(3): 93-99.
- Beyer, H.L. en P.A. Louw.** 2022. Aspekte van vernuwing in die Afrikaanse leksikografie: Standaard- en pedagogiese woordeboeke as barometer. *Lexikos* 32(3): 25-48.
- De Schryver, G.-M.** 2003. Lexicographers' Dreams in the Electronic-dictionary Age. *International Journal of Lexicography* 16(2): 143-199.
- Dobres, J., Chahine, N. en B. Reimer.** 2017. Effects of Ambient Illumination, Contrast Polarity, and Letter Size on Text Legibility under Glance-Like Reading. *Applied Ergonomics* 60: 68-73.
- Fuertes-Olivera, P.A. en S. Tarp.** 2014. *Theory and Practice of Specialised Online Dictionaries: Lexicography versus Terminography*. Berlyn: De Gruyter.
- Gouws, R.H.** 2001. The Use of an Improved Access Structure in Dictionaries. *Lexikos* 11: 101-111.
- Gouws, R.H.** 2005. Oor die verhouding tussen woordeboekstrukture, woordeboekinhoud en leksikografiese funksies. *Lexikos* 15: 52-69.
- Gouws, R.H.** 2017. Leksikografie. Carstens, W.A.M. en N. Bosman (Reds.). 2017. *Kontemporêre Afrikaanse Taalkunde*: 429-463. 2de uitgawe. Pretoria: Van Schaik.
- Gouws, R.H.** 2018a. Accessibility, Access Structures and Access Procedures. Jesenšek, V. en M. Enčeva (Reds.). 2018. *Wörterbuchstrukturen zwischen Theorie und Praxis*: 35-56. Berlyn: De Gruyter.
- Gouws, R.H.** 2018b. Internet Lexicography in the 21st Century. Engelberg, S., Kämper, H. en P. Storjohann (Reds.). 2018. *Wortschatz: Theorie, Empirie, Dokumentation*: 215-236. Berlyn: De Gruyter.
- Gouws, R.H.** 2018c. 'n Leksikografiese datatrekkingstruktuur vir aanlyn woordeboeke. *Lexikos* 28: 177-195.
- Gouws, R.H.** 2023. Nuwe raamstruktuurtypes in aanlyn woordeboeke. *Lexikos* 33(1): 382-403.
- Gouws, R.H. en T.J.D. Bothma.** 2024. Aspekte van inligtingsonttrekkingstrukture in aanlyn woordeboeke. *Lexikos* 34(1): 355-370.
- Gouws, R.H. en D.J. Prinsloo.** 2005. *Principles and Practice of South African Lexicography*. Stellenbosch: SUN PReSS.
- Gouws, R.H. en S. Tarp.** 2017. Information Overload and Data Overload in Lexicography. *International Journal of Lexicography* 30(4): 389-415.
- Gouws, R.H. en S. Tarp.** 2024. Despite Current Challenges: Lexicography Has a Bright Future. *Lexicographica* 40(1): 187-212.
- Index of Color Contrast Samples.** s.a.
<https://web.archive.org/web/20160814160227/>
<http://trace.wisc.edu/contrast-ratio-examples> [28 Februarie 2025]
- Jamil, A. en G. Denes.** 2024. Investigating Color-Blind User-Interface Accessibility via Simulated Interfaces. *Computers* 13(2): 53-68.
- Juicy Studio.** 2025. Luminosity Colour Contrast Ratio Analyser.
<https://juicystudio.com/services/luminositycontrastratio.php> [28 Februarie 2025]

-
- Kilgarriff, A. en I. Kosem.** 2012. Corpus Tools for Lexicographers. Granger, S. en M. Paquot (Reds.). 2012. *Electronic Lexicography*: 31-55. Oxford: Oxford University Press.
- Klosa, A. en R.H. Gouws.** 2015. Outer Features in e-Dictionaries. *Lexicographica* 31(1): 142-172.
- Mihindou, G.-R.** 2006. *A Theoretical Model for a Yipunu–English–French Explanatory Dictionary of Medical Terms*. Ongepubliseerde D.Litt.-verhandeling. Stellenbosch: Universiteit Stellenbosch.
- Müller-Spitzer, C. en A. Kopleinig.** 2014. Online Dictionaries: Expectations and Demands. Müller-Spitzer, C. (Red.). 2014. *Using Online Dictionaries*: 143-188. Berlyn: De Gruyter.
- Müller-Spitzer, C., Kopleinig, A. en A. Töpel.** 2012. Online Dictionary Use: Key Findings from an Empirical Research Project. Granger, S. en M. Paquot (Reds.). 2012. *Electronic Lexicography*. 425-457. Oxford: Oxford University Press.
- Rundell, M.** 2015. From Print to Digital: Implications for Dictionary Policy and Lexicographic Conventions. *Lexikos* 25: 301-322.
- Shieh, K.-K. en Y.-K. Lai.** 2008. Effects of Ambient Illumination, Luminance Contrast, and Stimulus Type on Subjective Preference of VDT Target and Background Color Combinations. *Perceptual and Motor Skills* 107(2): 336-352.
- Soegoto, E.S., M.A. Fadhlurrahman en Y.A. Hermawan.** 2019. Video Streaming Application Design for Color Blind Users. *Journal of Physics: Conference Series* 1402(6): 1-8.
- Swasty, W. en A.R. Adriyanto.** 2017. Does Color Matter on Web User Interface Design? *CommIT (Communication & Information Technology) Journal* 11(1): 17-24.
- Tarp, S.** 2012. Do We Need a (New) Theory of Lexicography? *Lexikos* 22: 321-332.
- TPGi.** 2025. Colour Contrast Analyser (CCA).
<https://www.tpgi.com/color-contrast-checker/> [28 Februarie 2025]
- WCAG (Web Content Accessibility Guidelines) 2.2.** 2024. *W3C Recommendation 12 December 2024*.
<https://www.w3.org/TR/WCAG21/> [27 Februarie 2025]
- WebAIM.** 2025. Contrast Checker.
<https://webaim.org/resources/contrastchecker/> [28 Februarie 2025]
- Wiegand, H.E.** 1983. Ansätze zu einer allgemeinen Theorie der Lexikographie. Schildt, J. en D. Viehweger (Reds.). 1983. *Die Lexikographie von heute und das Wörterbuch von morgen: Analysen — Probleme — Vorschläge*: 92-127. Berlyn: Akademie der Wissenschaften der DDR, Zentralinstitut für Sprachwissenschaft.
- Wiegand, H.E.** 1996. A Theory of Lexicographic Texts: An Overview. *South African Journal of Linguistics* 14(4): 134-149.
- Wiegand, H.E.** 1998. *Wörterbuchforschung: Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexicographie*. Berlyn: De Gruyter.
- Wiegand, H.E. en S. Beer.** 2013. Access Structures in Printed Dictionaries. Gouws, R.H., U. Heid, W. Schweickard en H.E. Wiegand (Reds.). 2013. *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*: 110-149. Berlyn/New York: De Gruyter.

Gender Bias in Computer-generated Thesauri: The Case of the Serbian Section of *Kontekst.io*, a Thesaurus of Synonyms and Semantically Related Terms

Dragana Čarapić, *Faculty of Philology, University of Montenegro, Nikšić, Montenegro* (draganac@ucg.ac.me)

(<https://orcid.org/0000-0003-3375-0169>)

and

Milica Vuković-Stamatović, *Faculty of Philology, University of Montenegro, Nikšić, Montenegro* (vmilica@ucg.ac.me)

(<https://orcid.org/0000-0001-5497-1361>)

Abstract: This paper studies gender bias in the computer-generated thesaurus *Kontekst.io*, which is a search portal of synonyms and semantically related terms in Serbian, Croatian and Slovenian. Its Serbian section, which is the focus here, is based on a natural language processing (NLP) technique called word embeddings and a large internet corpus of Serbian. Gender bias is uncovered in four selected entries of this thesaurus: *žena* (woman), *muškarac* (man), *d(j)evojka* (young woman) and *momak* (young man). The analysis is first conducted semantically and the terms found are grouped into various semantic fields. After that, in the vein of the earlier studies of gender bias in traditional dictionaries and critical discourse analysis, an analysis of gender bias in the selected entries is provided. The results show that gender bias is ubiquitous and that it extends deeper than the earlier studies of gender bias in word embeddings have shown. We then give recommendations for improving this lexicographic product based on the results.

Keywords: GENDER BIAS, COMPUTER-GENERATED THESAURUS, WORD EMBEDDINGS, *Kontekst.io*, SERBIAN, LEXICOGRAPHY

Opsomming: Geslagsvooroordeel in rekenaargegenereerde tesourusse: Die geval van die Serwiese afdeling van *Kontekst.io*, 'n tesourus van sinonieme en semanties verwante terme. In hierdie artikel word geslagsvooroordeel in die rekenaargegenereerde tesourus *Kontekst.io*, 'n soekportaal van sinonieme en semanties verwante terme in Serwies, Kroaties en Sloweens, bestudeer. Die Serwiese afdeling, waarop daar hier gefokus word, is gebaseer op 'n natuurliketaalprosesseringstegniek (NTP-tegniek) genaamd woordinbedding en 'n groot internetkorpus van Serwies. Geslagsvooroordeel word in vier uitgesoekte inskrywings in hierdie tesourus blootgelê: *žena* (vrou), *muškarac* (man), *d(j)evojka* (jong vrou) en *momak* (jong man). Die ontleding word eers semanties uitgevoer en die terme wat gevind word, word in verskillende semantiese velde ingedeel. Daarna, in dieselfde trant as vroeëre studies van geslagsvooroordeel in

tradisionele woordeboeke en kritiese diskoersanalise, word 'n ontleiding van geslagsvooroordeel in die uitgesoekte inskrywings verskaf. Die resultate toon dat geslagsvooroordeel alomteenwoordig is en dat dit verder strek as wat vroeëre studies van geslagsvooroordeel in woordinbedding aange-
toon het. Aanbevelings wat op die resultate gebaseer is, word dan gemaak om hierdie leksikogra-
fiese produk te verbeter.

Slutelwoorde: GESLAGSVOORoorDEEL, REKENAARGEGENEREEERDE TESOURUS, WOORDINBEDDING, *Kontekst.io*, SERWIES, LEKSIKOGRAFIE

1. Introduction

The recent rise of machine learning and artificial intelligence (AI) has transformed many fields. Among these, Natural Language Processing (NLP) has gained prominence by enabling computers to process and generate human language with ever increasing sophistication. Central to NLP are word embeddings — vector representations of words which capture their semantic meaning based on the context in which they are used in large corpora (Lee 2020). Word embeddings, however, reflect, perpetuate and even amplify societal biases from the source data, i.e. corpora, including gender bias (Gonen and Goldberg 2019). For instance, one type of word embeddings (*word2vec*), trained on the *Google News* dataset, in answer to the following: "man is to computer programmer as woman is to x", provides that "x is a homemaker" (Bolukbasi et al. 2016). Gender bias was demonstrated as omnipresent and consistent across different types of word embeddings and the proposed methods of mitigating it have so far produced only a limited effect — as Gonen and Goldberg put it, it is the effect of putting "a lipstick on a pig" (2019).

One area where word embeddings are starting to find its application is the creation of thesauri, especially for lesser-resourced languages, with poorly developed lexical databases (Arppe et al. 2023). Lexicographic products reflect but also shape our understanding of language and, by extension, the society (cf. Grenon-Nyenhuis 2000). Traditional dictionaries have been found to reflect and reproduce gender stereotypes (Bergenholtz and Gouws 2006; Nübling 2009). The same is true, and much more so, of word embeddings and the products based on them, given that they reflect gender bias existing in the source data. This fact should be taken into account when making lexicographic products based on word embeddings, bearing in mind that lexicographers have a moral responsibility to avoid reinforcing harmful gender stereotypes (Müller-Spitzer 2023).

This paper investigates gender bias in a computer-generated thesaurus intended for public use, aiming to uncover and point to some of such biases in it. The thesaurus in question is *Kontekst.io*, which contains semantically related words and synonyms in Serbian, Croatian and Slovenian. Our study will focus on the examples from Serbian. It should be noted that in this study we consider thesaurus as a type of a dictionary, following Gouws (2017: 134).

Gender bias in word embeddings has been studied relatively extensively so far (Zhao et al. 2019; Basta et al. 2019; Yang and Feng 2020; Basta et al. 2021;

Caliskan et al. 2022; etc.). However, as far as we know, the present study will be the first to offer a linguistic analysis of gender bias in a thesaurus based on word embeddings intended for general public use. In addition, the study of gender bias in word embeddings has mostly focused on the phenomenon of gender bias in occupation terms, while the present study will go beyond that — we will cover all synonyms and semantically related terms provided under several selected entries in Serbian (*žena* (woman), *muškarac* (man), *d(j)evojka* (young woman), and *momak* (young man)).

The current study builds on the body of research examining gender bias in traditional dictionaries and in word embeddings. The analysis will be performed within the vein of critical discourse analysis, given that we can understand dictionary as a text (cf. Fuertes-Olivera and Tarp 2022). In critical discourse analysis, ideology and power are key terms, and lexicography and its products are always subject to them, as they "are never value-free, apolitical or asocial" (Chen 2019: 1). Critical discourse analysts depart from the premise that language is not a mere mirror of social phenomena, but also their constitutive factor (Vuković-Stamatović 2022: 429) and that the analysis must simultaneously include the analysis of text, discourse practice, i.e. the origin, distribution and/or use of the text, and, finally, its social context (Fairclough 1992, 1995). Given that in this study we are dealing with a computer-generated thesaurus, we cannot criticise the lexicographer's intervention but we can criticise the lack thereof, as well as the data which are the source of bias. The approach to lexicography through critical discourse analysis belongs to critical lexicography in general (Kachru 1995; Chen 2019).

The background of the paper consists of two parts: Section 2 will cover the phenomenon of gender bias in traditional dictionaries, while Section 3 will present word embeddings and discuss them through the lens of gender bias. After that, in Section 4, we will present the thesaurus *Kontekst.io* and our method, following which the analysis is provided.

2. Gender bias in traditional dictionaries

In patriarchal societies, men are seen as the primary sources of power and moral authority. They also tend to hold prominent positions in decision-making, social privilege and property ownership. Certain cultures are far more patriarchal than others, even though most societies are patriarchal to some degree and assign gendered responsibilities (James 2010). This has traditionally had an impact, in particular, on women's status in these societies and has also been reflected in their languages. Dictionaries do not only reflect the sociolinguistic reality of certain speech communities — they also tend to propagate it, especially bearing in mind that they are frequently seen as authoritative sources which can shape how we understand words. As such, a dictionary can be "an agent of social impact" (Gouws 2022: 40).

Various research studies have highlighted the presence of gender bias in dictionaries. Some recent papers on the issue include: Norri 2019; Iversen 2021;

Solonets 2021; Pettini 2021; Müller-Spitzer and Rüdiger 2022; Vacalopoulou 2022; Fuertes-Olivera and Tarp 2022; Müller-Spitzer 2023, etc., indicating that gender bias in dictionaries is really not a thing of the past (cf. Fuertes-Olivera and Tarp 2022) and that modern dictionaries still feature it, despite the increasing efforts dictionary makers have been investing in reducing it.

Gender bias in dictionaries manifests in different ways. Some of these include asymmetrical definitions, stereotypical usage examples, and a choice of collocations which reflects stereotypes.

To illustrate asymmetrical definitions for different genders, we will use an example from the *Dictionary of the Serbian Language* (Matica Srpska 2007), where the entries *man* and *woman* are defined as follows:

- man – a person, human being of a male gender, a person who is of a gender opposite to woman (2007: 746),
- woman – 1. a human being who has the ability to give birth, of a gender opposite to man. 2. a person who is married, a wife. 3. a. an adult person of a female gender. b. a female person working in a house, a maid. 4. fig. pej. a weak person, a coward (when talking about men) • easy ~ a woman of low morale in her relation to men. *take a woman* – to get married (2007: 369).

The asymmetry in the two definitions above can be seen even physically — the entry related to *man* is much shorter to that for *woman*. *Man* is, first of all, a *person*, who does not need any more specific defining, whereas *woman* is firstly described through her role of giving birth (meaning 1). In addition, woman is stereotypically portrayed in meaning 2, where she is defined through her marital status; in meaning 3b, where she is defined through household chores and working for someone else; then in meaning 4, where a *woman* means a *weak* man; and in the phrase *easy woman*, with negative connotations, again in relation to men. The only neutral meaning for *woman* is, in fact, provided in 3a. The described gender difference is a result of the Serbian patriarchal society, which is reflected in the given dictionary entries. However, as said before, dictionaries do not just describe the sociolinguistic reality, but they propagate it (and, perhaps, co-construct it?), and one of the dilemmas before dictionary makers is how to intervene and to what extent.

As Müller-Spitzer (2023: 80) notes, lexicographers have a responsibility, given that dictionaries may contribute to exclusion and perpetuating gender stereotypes. In other words, "the representation of gender in dictionaries is a matter of both language use and lexicographic-moral responsibility" (Müller-Spitzer 2023: 83). Due to this, many dictionary makers have opted to invest effort into reducing gender bias as much as possible in their dictionaries. As a result, a stark asymmetry in the definitions of genders is more rarely seen in modern dictionaries, however, gender bias remains at more subtle levels. Müller-Spitzer and Rüdiger (2022: 130) analyse the choice of examples in the entries on *man*, *woman*, *boy* and *girl* in the *Cambridge Dictionary*, and quote the following examples from these entries:

"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."

"Who was that beautiful girl I saw you with last night?"

"Both girls compete for their father's attention."

The examples contain gender stereotypes — men and boys are associated with playing sports, drinking alcohol, and intelligence; women and girls are described in terms of appearance and as fighting for men's attention.

A similar situation may be noted with collocations. For instance, the *Online OXFORD Collocation Dictionary of English* (2019) (<https://m.freecollocation.com>) provides the following adjectival collocates for the entry *girl*:

- baby; little, small, young; adolescent, teenage; bubbly, happy, lively; lovely, nice; attractive, beautiful, good-looking, gorgeous, handsome, pretty, stunning; single, unmarried.

In contrast, consider the following adjectival collocates given by the same dictionary for the entry *boy*:

- big; little, small; young; elder, eldest, older; baby; adolescent, teenage; good; naughty; bright, clever.

As can be seen, some of the collocates for boys and girls are similar, especially those referring to age and size. However, considerable differences may be noted in other semantic fields — for instance, on the one hand, *girls* have six collocates for describing pleasant physical looks, four for describing pleasant disposition, and two for describing marital/partner status, as qualities considered important for females in anglophone societies, while boys are not described at all in terms of these three aspects. On the other hand, two collocates for boys refer to their intelligence, as a quality more associated with men in gender stereotypes, while such collocates are not found for girls. In addition, *boys* are described as *naughty*, also a quality more associated with men, as more obedience is expected from girls. The choice of the collocates certainly comes from the corpus that informs the *Online OXFORD Collocation Dictionary of English* and the same is true of the thesauri created from word embeddings.

In the literature, we have not come across examples treating gender bias reflected in a choice of synonyms (and semantically related terms) for certain entries, but, as our present analysis will show, gender bias can be uncovered through this lens as well.

3. Word embeddings and gender bias

As suggested earlier, a word embedding is a numeric representation of a word in the form of a vector. Such representations are used in Natural Language Pro-

cessing (NLP) to allow for automatic text analysis and processing to complete tasks such as sentiment analysis (analysis of texts to determine if their emotional tone is positive, neutral, or negative), translation, web search, parsing through different kinds of texts, etc. They have also been used to inform and make some dictionaries, thesauri in particular (Morinaga and Yamaguchi 2018; Chaimae et al. 2020; Liang et al. 2023; Arppe et al. 2023), although this use of word embeddings has not been as frequent as the uses mentioned above.

When using the technique called word embeddings, each word is assigned a number format, i.e. a real-valued vector. Words used in similar contexts, i.e. similarly, are closer in vector space — for example, the words "tomato" and "potato" would be close together in the vector space because they have certain similarities, e.g. they are both edible domesticated plants, while "tomato" and "house" would be further apart because they are used in largely different contexts. Thus, the techniques involving word embeddings measure and classify semantic similarities between words by examining their distributional properties within large language data.

For easier understanding of what word embeddings actually are, below we provide a graphic visualisation for some words in a simple, three-dimensional vector space:

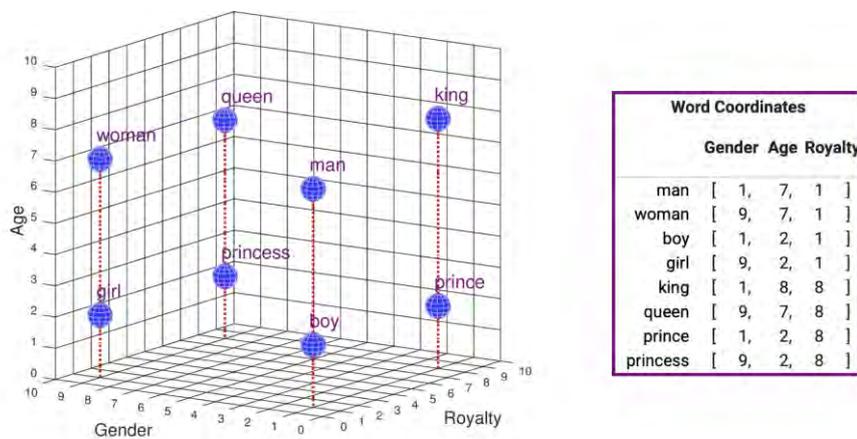


Figure 1: An illustration of word embeddings in a 3D vector space (Touretzky 2024)

In Figure 1, the words *boy*, *girl*, *man*, *woman*, *prince*, *princess*, *king* and *queen* are represented along three semantic dimensions: gender (smaller values indicate male gender, while larger indicate female), age (smaller values indicate younger age), and royalty (greater values for royal status). The values assigned are not perfectly symmetrical — for instance, for age, *king* is assigned an 8 while a queen is assigned a 7. The reason for this is that often in the corpora there is more talk

of *old kings* than *old queens*. Each of the coordinates in this 3D space can be read using three numbers, one for each of the three dimensions involved — the list of these three numbers or values for all the words depicted in the figure is given on the right. These series of numbers for each word represent their vectors. Of course, as we said, this 3D representation is very simplified and words, in practice, are represented by a much higher number of dimensions — of the order of tens and even hundreds of dimensions; in addition, the numbers are more precise and are rarely whole but rather decimal numbers. Based on the numbers assigned to each word, words can be mathematically compared in terms of their semantic similarity or semantic distance, and analogies such as "a man is to a woman as a king is to a queen" can be derived. Based on this methodology, it is possible to derive a list of synonyms and semantically related terms from a corpus, and as we have seen, these can be used to form lexical databases, inform dictionaries, and make thesauri.

As suggested earlier, gender bias is rather frequent in word embeddings, primarily as a reflection of the source data which naturally contain it. Thus, word embeddings can yield analogies such as "a father is to a doctor as a mother is to a nurse" (Bolukbasi et al. 2016). However, word embeddings do not just reflect the source data and the biases in them; the algorithms which they are based on can even inadvertently amplify them (Bolukbasi et al. 2016; Basta et al. 2019). In response to this problem, different debiasing methods have been proposed — some of these focus on correcting the source data itself, some intervene in the algorithm, while some use human-made dictionaries for creating less biased word embeddings (Zhao et al. 2019; Yang and Feng 2020, etc.). However, the results are mixed and what is certain is that gender bias always remains in the output (Gonen and Goldberg 2019; Lee 2020). Most debiasing approaches focus on the words related to occupations, but gender bias extends much deeper and can be hidden across multiple features (Caliskan et al. 2022). So far, the debiasing methods have been successful at rather "hiding gender bias but not truly removing it", Ronchieri and Biagi (2023: 730) conclude.

4. Data and method

As suggested earlier, the data used in this study derives from the computer-generated thesaurus *Kontekst.io*. In the following text, we present *Kontekst.io* in more details, following which more light is shed on the method.

4.1 *Kontekst.io*

Kontekst.io is a search portal of synonyms and semantically related terms in Serbian, Croatian and Slovenian, based on word embeddings (Plahuta 2024). On its website, it is stated that it can primarily be used as "a dictionary of synonyms".

What the homepage of *Kontekst.io* looks like can be seen in Figure 2 below:



Figure 2: The homepage of *Kontekst.io*

As can be seen, the website's homepage is simple — the user first chooses one out of the three languages offered. The second step is to type in the search term or to choose one of the popular search terms offered below.

The output offers a list of words with similar distributional properties, i.e. semantically related words including synonyms and even antonyms, along with the normalised frequency of the word per 1 million words of the corpus, its score of semantic similarity with the entered term and the examples of its use in the source corpus. Below is part of the output for the entry *lijepa* (beautiful) from the Serbian section (Figure 3); we present only part of it, as the whole output contains some 50 synonyms and semantically related terms and cannot fit here.

Kontekst.io
slični izrazi i sinonimi u savremenom srpskom, hrvatskom i slovenskom
SL HR SR

Sličnost reči ili fraza u rezultatima zavisi od toga, koliko puta se reč ili fraza pojavlja u sličnom kontekstu kao "lijepa".

Slični izrazi i sinonimi za	UČESTALOST	SLIČNOST	Primeri iz općenitog korpusa
lijepa			Korpus srWac
Kliknite za traženje			srWac je korpus srpskog jezika (Charva.it)
lepa	36.67	82%	moj mentor i ja, dijelili smo međusobno poštovanje. I iz poštovanja, izraslo je prijateljstvo, i jedna lijepa
prekrasna	1.69	80%	ničeg nejasnog i mračnog. Svoj život on je proveo u mjestima koja su uvijek bila u prirodi i vanredno lijepa
prelijepa	0.76	80%	tri zvezde imagjaše''. Jednog jutra, Veljko tjera u polje, da pasu, veliku svoja krmaču i tri vrlo lijepa
predivna	4.46	78%	u skladu sa povodom koncert je trebalo da poste 50 minuta bude zaključ? en legendarnom " Bila je tako lijepa
divna	11.47	78%	
krasna	0.67	77%	
zgodna	5.57	77%	

Figure 3: Some synonyms and semantically related terms for *lijepa* (beautiful) in *Kontekst.io* (Serbian)

The Serbian thesaurus is based on the Serbian web corpus *srWaC* (Ljubešić and Klubička 2016), which contains 554,627,647 running words (version 1.1), and was obtained by crawling the *.rs* web domain (the Serbian national web domain). Some of the advantages of using this corpus include its size, the fact that it is quite recent and the fact that it is available for public use. Additionally, there are not many viable corpus alternatives when it comes to corpora for Serbian, which is far less resourced than many other European languages. The disadvantage of the corpus is that all the texts come from the Internet.

As we find in Ulčar et al. (2021), *Kontekst.io* was made using the *word2vec* algorithm (Mikolov et al. 2013) and a 256-dimensional vector model trained for its needs. Using their mathematical approach, Ulčar et al. (2021) have also determined that *Kontekst.io* contains a high degree of gender bias in its entries referring to occupations, but they have not looked beyond that.

According to the data from *Similarweb.com*, a website tracking 100 million websites in 190 countries, *Kontekst.io* was the ninth most popular website in the category of dictionaries and encyclopedias in Serbia in April 2024. Most of the websites preceding it were global websites, such as *Wikipedia.org*. Based on this, we can say that *Kontekst.io* is a rather popular thesaurus in Serbia. Most of its traffic comes from the referrals from the *Google* search when users are looking for synonyms for certain words (again, according to *Similarweb.com*).

4.2 Method

In this study, we analyse the synonyms and semantically related terms provided under four selected entries: *žena* (woman), *muškarac* (man), *d(j)evojka* (young woman), and *momak* (young man), in the Serbian component of *Kontekst.io*. The number of the entries studied was limited by the space provided for this paper. The entries chosen are obviously gendered words, which is why it was expected that their analysis could most clearly point to the presence and types of gender bias in *Kontekst.io*.

To uncover and analyse gender bias in the selected entries, we use a method which encompasses comparative analysis, grouping the terms in separate semantic fields and applying the connotative analysis of the specific terms. The method embodies different levels of paradigmatic relations, starting from the hyponymy — superordinate term, and continuing with the synonymous expressions allocated to the respective semantic fields. For better visualisation, the results thus processed are presented in tables, comparatively for *woman* and *man*, as well as for *young woman* and *young man*. This level of analysis is purely semantic.

Further, the semantic method is complemented with critical discourse analysis (Fairclough 1992, 1995), given that the ideology underlying the selection of the terms presented as synonyms and semantically related terms, is uncovered and discussed. The selection is computer-produced but it reflects the bias existing in the source data which feed the thesaurus. More specifically, the analysis of the entries is in the vein of the analyses of gender bias in traditional dictionaries

presented in Section 2: our own examples and those of Müller-Spitzer (2023). Thus, the inequalities in the choice and the frequency of the different synonyms and semantically related terms (provided for the said male and female entries) are noted and discussed.

5. Results and analysis

Given that two pairs of entries are analysed in this paper, the analysis will be organised in two parts, each dedicated to one of these.

5.1 The entries for *žena* (woman) and *muškarac* (man) in Serbian *Kontekst.io*

The terms provided as synonyms and semantically related terms for the entries *woman* and *man* have first been grouped into separate semantic fields, as explained in the Method subsection. The results for Serbian *Kontekst.io* are presented in Table 1 below.

Table 1: Synonyms and semantically related terms for the entries *žena* (woman) and *muškarac* (man) in Serbian in *Kontekst.io*¹

<i>Kontekst.io</i> (SERBIAN)	
ŽENA (WOMAN)	MUŠKARAC (MAN)
SUPERORDINATE TERM: Osoba (person)	SUPERORDINATE TERM: Čovek, čovjek (human, man), stvor (creature)
AGE-RELATED: Beba (baby), djevojčica, djevojčica (little girl), djevojka, devojka, devojaka (young woman), cura (lass), curica (lassy DIM. ²), mlada dama (young lady), mlada devojka (young girl), starica (old woman), bakica (granny)	AGE-RELATED: Mališan (tot), dečak, dječak, dečko (boy), klinac (kid), mladić, momak (young man), tinejdžer (teenager), adolescent (adolescent), pedesetogodišnjak (fifty-year-old man), matorac (old man)
OCCUPATION-RELATED: Domaćica (housewife), dadilja (nanny), sluškinja (maid)	OCCUPATION-RELATED: Policajac (police officer), vojnik (soldier)
FAMILY-RELATED: Trudnica (pregnant woman), majka (mother), udovica (widow), porodica (family)	FAMILY-RELATED: Muž (husband), roditelj (parent)
ANIMAL-RELATED: Kuja (bitch PEJ. ³)	ANIMAL-RELATED: Mužjak (male), pas (dog), bik (bull), majmun (monkey), konj (horse)
EVIL-RELATED: Veštica (witch PEJ.)	EVIL-RELATED: Nasilnik (bully), vampire (vampire)

PATIENT/BENEFICIARY ROLES: Pacijentkinja (patient FEM. ⁴), pacijentica (patient FEM.), zatvorenica (prisoner FEM.), mušterija (customer FEM.), klijentkinja (client FEM.)	PATIENT/BENEFICIARY ROLES: Pacijent (patient)
APPEARANCE-RELATED: Plavuša (blonde), crnkinja (black woman), belkinja (white woman), ženica (woman DIM.)	APPEARANCE-RELATED: Belac, bijelac (white man), crnac (black man)
FOREIGNER: Francuskinja (French woman)	FOREIGNER: Stranac (foreigner), Indijanac (Indian), Amerikanac (American)
MONEY-RELATED: Bogatašica (rich woman)	MONEY-RELATED: Bogataš (rich man), beskućnik (homeless man)
SOCIAL ROLE-RELATED: Dama (lady), komšinica (neighbour FEM.), prijateljica, drugarica (friend FEM.)	SOCIAL ROLE-RELATED: Džentlmen (gentleman)
OTHER: Duša (darling), seljančica (village girl)	SEXUALITY-RELATED: Frajer (hot shot), ljubavnik (lover)

The first thing which we can notice in Table 1 is that the superordinate term for *muškarac* (man) is *čov(j)ek* (human), whereas for *žena* (woman) it is *osoba* (person). Further, the following can also be noted (the observations are presented according to the semantic fields):

- **age:** more diminutives and the adjective *young* preceding the nouns designating women, e.g. *djevojka*, *devojka*, *devojaka* (young woman), *mlada dama* (young lady), *mlada devojka* (young girl), *curica* (lassy DIM.);
- **occupation:** gender bias detected under both the entries, e.g. *domaćica* (housewife), *dadilja* (nanny), and *sluškinja* (maid) under the entry *woman*, and *policajac* (police officer) and *vojniki* (soldier), under the entry *man*;
- **family:** there are more terms in this field for *woman* than *man*, e.g. *trudnica* (pregnant woman), *majka* (mother), *udovica* (widow), *porodica* (family), are all used for *woman*, while only *muž* (husband) and *roditelj* (parent) are used for *man*.
- **animal:** only one term with negative connotations is used for *woman*: *kuja* (bitch PEJ.); however, there are five terms in this field for *man*: *mužjak* (male), *pas* (dog), *bik* (bull), *majmun* (monkey), and *konj* (horse), some of which have negative connotations, too;
- **evil:** gender bias detected in the derogatory terms *veštica* (witch PEJ.) for *woman*, and *nasilnik* (bully) and *vampire* (vampire) for *man*;
- more **appearance**-related terms for women: *plavuša* (blonde), *crnkinja* (black woman), *belkinja* (white woman), *ženica* (woman DIM.);

- more **socially-** and **family-**defined **roles** for *woman*: *dama* (lady), *komšinica* (neighbour FEM.), *prijateljica*, *drugarica* (friend FEM.);
- more **money-**related characterisations for men, e.g. two depicting *man* (*bogataš* (rich man), and *beskućnik* (homeless man)) vs. one used for *woman* (*bogatašica* (rich woman));
- **sexuality-**related terms — only used for men, whereby both the terms in this field have a positive meaning: *frajer* (hot shot), *ljubavnik* (lover);
- **other terms** for *woman*: there is one endearing term in the data and it is used for *woman* (*duša* (darling)); the other term in this category is *seljančica* (village girl) and it is condescending.

We proceed with the results for the second pair of entries analysed.

5.2 The entries for *d(j)evojka* (young woman) and *momak* (young man) in Serbian *Kontekst.io*

In the same vein, we conduct the analysis for the entries: *d(j)evojka* (young woman) and *momak* (young man). The results are given in Table 2.

Table 2: Synonyms and semantically related terms for the entries *d(j)evojka* (young woman) and *momak* (young man) in Serbian in *Kontekst.io*

<i>Kontekst.io</i> (SERBIAN)	
D(J)EVOJKA (YOUNG WOMAN)	MOMAK (YOUNG MAN)
SUPERORDINATE TERM: Dama (lady)	SUPERORDINATE TERM: Čova, čovek, čovjek (human, man), čovečuljak (man DIM.), muškarac (man)
AGE-RELATED: Bakica (granny DIM.), cura, curica (girl), devojčica, djevojčica, klinka, mala devojčica (little girl), mlada devojka (young girl), starica (old woman), tinejdžerka (teenage girl)	AGE-RELATED: Čikica (old man DIM.), dečak, dečkić (young boy DIM.), dečko, dječak, deran, klinac (young boy), matorac (old guy), mladić, momčić (young man DIM.)
OCCUPATION-RELATED: Dadilja (nanny), konobarica (waitress), služavka (maid), stažistica (intern FEM.)	OCCUPATION-RELATED: Pandur (cop), policajac (police officer)
FAMILY-RELATED: Majka (mother), sestra (sister)	FAMILY-RELATED: Muž (husband), roditelj (parent)
ANIMAL-RELATED: Kučka (bitch), mačka (cat)	ANIMAL-RELATED: Džukac (mutt), majmun (monkey), pas (dog)

EVIL-RELATED: Ludača (mad woman)	EVIL-RELATED: Gad (bastard), govнар (piece of shit DER. ⁵), kučkin sin, (son of a bitch DER.), kurvin sin (son of a whore DER.), ludak (mad man), seronja (asshole DER., SLANG), tip (guy)
APPEARANCE-RELATED: Lepa devojka (pretty girl), lepotica (beauty), plavuša (blonde)	APPEARANCE-RELATED: Crnac (black man), debeljko (fat man), lepotan (handsome man), plavušan (blonde man)
SOCIAL ROLE-RELATED: Cimerka (roommate FEM.), drugarica (friend FEM.), komšinica (neighbour FEM.), nevesta (bride), poznanica (acquaintance FEM.), prijateljica (friend FEM.), verenica (fiancé FEM.)	SOCIAL ROLE-RELATED: Drugar (friend), dasa (hot shot), klipан (loon DER.), ortak (buddy)
SEXUALITY-RELATED: Fufa, fufica, kurva (hooker), prostitutka (prostitute FEM.)	SEXUALITY-RELATED: Frajer (hot shot), jebač (fucker DER., SLANG)
MONEY-RELATED: Bogatašica (rich woman)	FOREIGNER: Indijanac (Indian), Amerikanac (American)

The following can be noted regarding the results in Table 2:

- **age:** more diminutive expressions are used preceding the noun for *young man*: *čovečuljak* (man DIM.), *čikica* (old man DIM.), *dečkić* (young boy DIM.), *momčić* (young man DIM.);
- **occupation:** gender bias is found under both *young woman* and *young man*, e.g. *dadilja* (nanny), *konobarica* (waitress), *služavka* (maid), *stažistica* (intern FEM.), in the former category, and *pandur* (cop) and *policaјac* (police officer) in the latter;
- **family:** the same number of synonyms/semantically related terms are provided for both *young woman* and *young man*: *majka* (mother) and *sestra* (sister), vs. *muž* (husband) and *roditelj* (parent);
- **animal:** on the one hand, two terms are used for *young woman*: *kučka* (bitch), which has negative connotations, and *mačka* (cat), which has sexual connotations; on the other hand, three terms are used with negative connotations for *young man*: *džukac* (mutt), *majmun* (monkey), and *pas* (dog);
- **evil:** in this semantic field, there is one term for *young woman* (*ludača* (mad woman)), whereas there are seven derogatory terms for *young man*: *gad* (bastard DER.), *govнар* (piece of shit DER.), *kučkin sin* (son of a bitch DER.), *kurvin sin* (son of a whore, DER.), *ludak* (mad man), *seronja* (asshole DER., SLANG), and *tip* (guy);

-
- more **appearance**-related terms are used for *young man*, e.g. *crnac* (black man), *debeljko* (fat man), *lepotan* (handsome man), and *plavušan* (blonde man), than for *young woman*: *lepa devojka* (pretty girl), *lepotica* (beauty), and *plavuša* (blonde);
 - there more **socially defined roles** under the entry *young woman*: *cimerka* (roommate FEM.), *drugarica* (friend FEM.), *komšinica* (neighbour FEM.), *nevesta* (bride), *poznatica* (acquaintance FEM.), *prijateljica* (friend FEM.), *verenica* (fiancé), than for *man*: *mladoženja* (bachelor), *razbojnik* (scourer), and *zatvorenik* (prisoner);
 - there is one **money**-related characterisation for *young woman*: *bogatašica* (rich woman), and there are no such characterisations for *young man*;
 - there are two expressions for **foreigner** under the entry *young man*: *Indijanac* (Indian) and *Amerikanac* (American);
 - **sexuality**: four terms from this field are used for *young woman*: *fufa*, *fufica*, *kurva* (hooker) and *prostitutka* (prostitute FEM.), while two such terms are used for *young man*: *frajer* (hot shot) and *jebač* (fucker DER., SLANG).

A discussion of these findings is presented in Section 6.

6. Discussion

In the analysis of the selected terms in *Kontekst.io* in Serbian, several elements of gender bias have been detected.

On the one hand, the superordinate term for both *muškarac* (man) and *momak* (young man) is *čov(j)ek* (human), whereas for *žena* (woman) it is *osoba* (person) and for *d(j)evojka* (young woman) it is *dama* (lady). The male terms are thus *more encompassing*, given that they can stand for the entire human kind, i.e. both men and women, unlike the female terms, which are much more specific and exclusive in the sense that they cannot extend their meaning beyond the female sex. Even more specifically, *person* stands for just one individual. Additionally, *lady* is a term which reflects a socially preferable role for women.

Another finding is that there are more diminutives under the entries for *woman*, as well as that the adjective *young* often precedes the nouns in this category. This has to do with the general notion of woman as a weaker sex in a patriarchal society. As such, woman requires more affection and protection through the use of diminutives.

We have also noticed that there is gender bias under all the analysed entries related to occupations. As said earlier, this phenomenon has been investigated earlier (for instance, Ulčar 2021) and it has been shown that there is substantial gender bias in word embeddings when it comes to the names of occupations.

In *Kontekst.io*, more family-related terms are provided under the entry *woman*

than *man*. This result is reflective of the social position of woman, who is largely defined through the role she has within her family in a patriarchal society. Furthermore, it is noteworthy that there is no mention of *father* as a semantically related term for either *man* or *young man*, whereas *mother* is identified as a semantically related term for both *woman* and *young woman*. The only family-related word that is provided for *man* and *young man* is *parent*, a gender-neutral designation that does not specify paternal identity. Such a disbalance is not seen, however, in the second pair of the entries, given that the additional family roles for females are assumed only after a certain age.

There are more terms in the semantic field *animal* for *man* and *young man*, than under the female entries. Many of these terms used for men have negative connotations, as they are associated with negative character traits: being *aggressive*, *violent* and *fierce*. This is also the result of gender bias existing in society and, consequently, in the source data used for the thesaurus.

Likewise, in the semantic field relating to *evil*, there are many more terms for *man* than *woman*. This result points to a similar notion of *man* as a sex more capable of inflicting violence and other evil acts and it also reveals that it is socially more acceptable to address other men rather than women using such terms. Further, of the two terms supplied in this field for women, *witch* particularly stands out and also reflects gender bias.

More appearance-related terms for the entry *woman* show that woman is more socially defined by her physical appearance than man. The terms relating to *young woman* and *young man*, however, are similar in number, suggesting that appearance becomes much more relevant for women only after a certain age, the same was noted for socially defined roles for women.

More money-related characterisations for men imply the importance of *money* in defining *man*. Also, in our data, foreigners tend to be men more often than women.

There are some more sexual terms for *young woman* than *young man*, but it should be noted that all the terms from this field have negative connotations. All of them are *derogatory* slang expressions and they imply negative character traits.

As we can see, gender bias is ubiquitous in all the entries studied. This is not just gender bias which is related to the female gender, which is usually more obvious — on the contrary, in the terms under the said entries we find reflections of many stereotypical characterisations of men, too. Thus, men are seen as being aggressive and violent, money is important for how they are socially defined, and they are seen to be working in certain professions, such as being a policeman or a soldier.

Some of the gender bias described here is more "innocuous" than the other. Perhaps on the more "innocuous" end of the spectrum are the terms for occupations, which have been the most frequent topic of interest for those seeking to address gender bias in word embeddings through improving their algorithms (Section 2). We believe that it is much more serious when the terms such

as *witch*, *prostitute*, *bitch*, and *mad woman*, for example, are provided as synonyms or semantically-related terms for *woman* or *young woman*. Most of the users of *Kontekst.io* will not be familiar with the discipline of semantics or the meaning of "semantically-related", especially in terms of semantic distribution, while many will understand what synonyms are and will perhaps see the terms provided as such. As suggested earlier, the website of *Kontekst.io* itself states that it can primarily be used as "a dictionary of synonyms", while most of its traffic comes from the *Google* search for synonyms. That being said, it does not help that this search portal does not contain instructions or a help page providing an explanation for common users on what these terms mean and what the results provided by the thesaurus actually show. We recommend that such guidance be provided visibly for the users.

Moreover, we have noticed that for derogatory terms and other terms with negative connotations, which were quite frequent in the results, no such designations were provided in order to point to users that these terms are not in common, everyday use. Providing the necessary flags, as is customary in dictionaries, should be the next step in improving *Kontekst.io*. Perhaps these can be imported from other dictionaries.

Additionally, it goes without saying that all products, including *Kontekst.io*, which are based on word embeddings, must be further improved towards eliminating more gender bias, whereby the process should not solely or primarily focus on the occupation terms. Lexicographic moral responsibility is not absolved if the thesaurus is created by a machine, i.e. an algorithm — people are always behind those algorithms, in the same way that they are behind the corpora informing the machines.

Gender bias can probably never be fully eliminated in any dictionary in general and in computer-generated ones in particular — we have demonstrated here how deep it can go in many different directions and for both the genders here studied. However, when making lexicographic products which are publicly available, one must be aware of the harms which such products may do, as they do not only reflect the existing gender bias but also may perpetuate and co-construct it.

7. Conclusion

This paper investigated gender bias in the computer-generated thesaurus *Kontekst.io*, aiming to uncover examples of this bias in it and give recommendations for improving it as a lexicographic product. *Kontekst.io* is a search portal for synonyms and semantically related terms in Serbian, Croatian and Slovenian, and this paper focused on its Serbian section — specifically, the terms provided under the entries: *žena* (woman), *muškarac* (man), *d(j)evojka* (young woman), and *momak* (young man).

Gender bias in the studied entries was found to be ubiquitous and to run deeper than the earlier studies of this phenomenon in word embeddings have

shown. Earlier studies were mostly conducted by non-linguists who focused on the more obvious examples of gender bias and concentrated more on improving the algorithms than on detecting all types of gender bias in their products.

Based on the results, we gave some recommendations how *Kontekst.io* (and other similar lexicographic products) may be improved. These include providing clear definitions and instructions to users explaining what kind of terms are supplied in the entries, providing flags marking the use of certain words as derogatory, slang etc., and working further on the algorithms, thus eliminating more gender bias, especially beyond the occupation terms.

Ultimately, we believe that gender bias cannot be reduced to a certain satisfactory level without human intervention in lexicographic products such as the one studied here. However, we do understand the necessity of such products in lesser-resourced languages such as Serbian, but we hope that they can be additionally improved and the users more instructed into what they really are and what their limitations are, so as to avoid harmful consequences such as perpetuating and co-constructing gender bias.

Our analysis focused on the entries where gender bias might be more salient. Further investigation should focus on more entries and other language sections of *Kontekst.io* in order to uncover the more subtle forms of gender bias in them and produce further recommendations for improving this and other similar lexicographic products.

Endnotes

1. Note: Mistyped words have been excluded here.
2. Diminutive.
3. Pejorative.
4. Feminine.
5. Derogatory.

References

- Arppe, A., A. Neitsch, D. Dacanay, J. Poulin, D. Hieber and A. Harrigan. 2023. Finding Words that Aren't There: Using Word Embeddings to Improve Dictionary Search for Low-resource Languages. *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*: 144-155. Toronto, Canada: Association for Computational Linguistics.
- Basta, C., M.R. Costa-jussà and N. Casas. 2019. Evaluating the Underlying Gender Bias in Contextualized Word Embeddings. *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*: 33-39. Florence, Italy: Association for Computational Linguistics.
- Basta, C., M.R. Costa-jussà and N. Casas. 2021. Extensive Study on the Underlying Gender Bias in Contextualized Word Embeddings. *Neural Computing and Applications* 33(8): 3371-3384.
- Bergenholtz, H. and R. Gouws. 2006. How to Do Language Policy with Dictionaries. *Lexikos* 16: 13-45.

- Bolukbasi, T., K.-W. Chang, J.Y. Zou, V. Saligrama and A.T. Kalai.** 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. Lee, D., M. Sugiyama, U. Luxburg and I. Guyon and R. Garnett. 2016. *Advances in Neural Information Processing Systems 29. (30th Annual Conference on Neural Information Processing Systems 2016 (NIPS 2016), 5-10 December 2016, Barcelona, Spain)*: 4349-4357. La Jolla, CA, USA: NIPS Foundation.
- Caliskan, A., P.P. Ajay, T. Charlesworth, R. Wolfe and M.R. Banaji.** 2022. Gender Bias in Word Embeddings: A Comprehensive Analysis of Frequency, Syntax, and Semantics. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society (AIES 2022)*: 156-170. Oxford: AIES.
- Chaimae, A., M. Rybinski, E.Y. Yacine and J.F. Aldana Montes.** 2020. Comparative Study of Arabic Word Embeddings: Evaluation and Application. *International Journal of Computer Information Systems and Industrial Management Applications* 12: 349-362.
<https://cspub-ijcisim.org/index.php/ijcisim/article/view/469>
- Chen, W.** 2019. Towards a Discourse Approach to Critical Lexicography. *International Journal of Lexicography* 32(3): 362-388.
- Fairclough, N.** 1992. *Discourse and Social Change*. Cambridge: Polity.
- Fairclough, N.** 1995. *Critical Discourse Analysis*. London/New York: Longman.
- Fuertes-Olivera, P.A. and S. Tarp.** 2022. Critical Lexicography at Work: Reflections and Proposals for Eliminating Gender Bias in General Dictionaries of Spanish. *Lexikos* 32(2): 105-132.
- Gonen, H. and Y. Goldberg.** 2019. Lipstick on a Pig: Debiasing Methods Cover up Systematic Gender Biases in Word Embeddings but Do not Remove Them.
arXiv preprint arXiv:1903.03862
- Gouws, R.H.** 2017. Van tematiese na alfabetiese na tematiese ordening in woordeboeke — wisselwerking tussen teorie en praktyk. *Stellenbosch Papers in Linguistics Plus* 53: 133-148.
- Gouws, R.H.** 2022. Dictionaries as Instruments of Exclusion and Inclusion: Some South African Dictionaries as Case in Point. *Lexicographica* 38(1): 39-61.
<https://doi.org/10.1515/lex-2022-0003>
- Grenon-Nyenhuis, C.** 2000. The Dictionary as a Cultural Institution. *Intercultural Communication Studies* 10(1): 159-166.
- Iversen, S.H.** 2021. The (Re)presentation of Knowledge about Gender in Children's Picture Dictionaries. Goga, N., S.H. Iversen and A.-S. Teigland (Eds.). 2021. *Verbal and Visual Strategies in Nonfiction Picturebooks: Theoretical and Analytical Approaches*: 67-79. Oslo: Scandinavian University Press.
- James, K.** 2010. Domestic Violence within Refugee Families: Intersecting Patriarchal Culture and the Refugee Experience. *Australian and New Zealand Journal of Family Therapy* 31(3): 275-284.
- Kachru, B.B.** 1995. Afterword: Directions and Challenges. Kachru, B. B. and H. Kahane (Eds.). 1995. *Cultures, Ideologies, and the Dictionary: Studies in Honor of Ladislav Zgusta*: 417-424. Tübingen: Max Niemeyer.
- Lee, E.** 2020. Gender Bias in Dictionary-derived Word Embeddings. Technical Report (CS230: Deep Learning, Fall 2020, Stanford University, CA).
http://cs230.stanford.edu/projects_fall_2020/reports/55476615.pdf
- Liang, H., Y.M.M. Ng and N.L. Tsang.** 2023. Word Embedding Enrichment for Dictionary Construction: An Example of Incivility in Cantonese. *Computational Communication Research* 5(1): 1-26.
- Ljubešić, N. and F. Klubička.** 2016. *Serbian Web Corpus srWaC 1.1*, Slovenian Language Resource repository CLARIN.SI, ISSN 2820-4042.
<http://hdl.handle.net/11356/1063>.

- Mikolov, T., I. Sutskever, K. Chen, G. Corrado and J. Dean.** 2013. Distributed Representations of Words and Phrases and their Compositionality. *arXiv preprint arXiv:1310.4546*
- Morinaga, Y. and K. Yamaguchi.** 2018. Improvement of Reverse Dictionary by Tuning Word Vectors and Category Inference. Robertas Damaševičius, R. and G. Vasiljeviene (Eds.). 2018. *Information and Software Technologies: 24th International Conference, ICIST 2018, Vilnius, Lithuania, October 4–6, 2018, Proceedings 24*: 533-545. New York: Springer.
- Müller-Spitzer, C.** 2023. Gender Stereotypes in Dictionaries: The Challenge of Reconciling Usage-based Lexicography with the Role of Dictionaries as Social Agents. *Lexikos* 33(2): 79-94.
- Müller-Spitzer, C. and J.O. Rüdiger.** 2022. The Influence of the Corpus on the Representation of Gender Stereotypes in the Dictionary. A Case Study of Corpus-based Dictionaries of German. Klosa-Kückelhaus, A., S. Engelberg, C. Möhrs and P. Storjohann (Eds.). 2022. *Dictionaries and Society. Proceedings of the XX EURALEX International Congress*: 129-141. Mannheim: IDS-Verlag.
- Norri, J.** 2019. Gender in Dictionary Definitions: A Comparison of Five Learner's Dictionaries and their Different Editions. *English Studies* 100(7): 866-890.
- Nübling, D.** 2009. Zur lexikografischen Inszenierung von Geschlecht. Ein Streifzug durch die Einträge von Frau und Mann in neueren Wörterbüchern. *Zeitschrift für germanistische Linguistik* 37(3): 593-633.
<https://doi.org/10.1515/ZGL.2009.037>
- Online OXFORD Collocation Dictionary of English.** 2019.
<https://m.freecollocation.com>
- Pettini, S.** 2021. One is a Woman, so That's Encouraging too. The Representation of Social Gender in "Powered by Oxford" Online Lexicography. *Lingue e Linguaggi* 44: 275-295.
- Plahuta, M.** 2024. *Kontekst.io*.
<https://virostatiq.com/> [Accessed 20 July 2024]
- Ronchieri, E. and C. Biagi.** 2023. Comparing Methods for Mitigating Gender Bias in Word Embedding. Bui, T.X. (Ed.). 2023. *Proceedings of the 56th Hawaii International Conference on System Sciences, January 3–6, 2023*: 722-731. Honolulu: HICSS.
<https://hdl.handle.net/10125/102720>
- Similarweb.** 2024. <https://www.similarweb.com/website/blog.context.io/> [Accessed 10 July 2024]
- Solonets, P.V.** 2021. *Gender and Dictionary: Russian Perspective*. Unpublished Master's Thesis. Braga: University of Minho.
- Touretzky, D.** 2024. Word Embedding Demo (webpage), Carnegie Mellon University blogs.
<https://www.cs.cmu.edu/~dst/WordEmbeddingDemo/tutorial.html>
- Ulčar, M., A. Supej, M. Robnik-Šikonja and S. Pollak.** 2021. Slovene and Croatian Word Embeddings in Terms of Gender Occupational Analogies. *Slovenščina 2.0: empirične, aplikativne in interdisciplinarne raziskave* 9(1): 26-59.
- Vacalopoulou, A.** 2022. Gender Stereotypes in Greek Children's Dictionaries. *Dictionaries* 43(1): 167-192.
- Vujanić, M., D. Gortan-Premk, M. Dešić, R. Dragičević, M. Nikolić, LJ. Nogo, V. Pavković, M. Radović-Tešić, N. Ramić, R. Stijović and E. Fekete.** 2007. *Rečnik srpskog jezika* [Dictionary of the Serbian Language]. Novi Sad: Matica Srpska.
- Vuković-Stamatović, M.** 2022. "Accessing the EU Is Like Running on a Treadmill in the Gym": How the EU Accession Process is Metaphorically Presented in the Online Media of Serbia, Montenegro, and Bosnia and Herzegovina. *Annales. Series historia et sociologia* 32(3): 427-448.

<http://lexikos.journals.ac.za>; <https://doi.org/10.5788/35-2-2108> (Volle uitgawe / Full volume)
<https://lexikos.journals.ac.za>; <https://doi.org/10.5788/35-2-2068> (Article)

- Yang, Z. and J. Feng.** 2020. A Causal Inference Method for Reducing Gender Bias in Word Embedding Relations. *Proceedings of the AAAI Conference on Artificial Intelligence* 34(5): 9434-9441.
<https://ojs.aaai.org/index.php/AAAI/article/view/6486>
- Zhao, J., T. Wang, M. Yatskar, R. Cotterell, V. Ordonez and K.-W. Chang.** 2019. Gender Bias in Contextualized Word Embeddings.
arXiv preprint arXiv:1904.03310

'n Omvattender aanbod van semantiese data in aanlyn woordeboeke*

Rufus H. Gouws, *Departement Afrikaans en Nederlands,
Universiteit Stellenbosch, Suid-Afrika*
(rhg@sun.ac.za)
(<https://orcid.org/0000-0002-3423-058X>)

Opsomming: Betekenis is die datatipe waaraan die meeste gebruikers van verklarende woordeboeke 'n behoefte het. Weens ruimtebeperkings kan gedrukte woordeboeke slegs in 'n beperkte mate verslag doen van betekenisaspekte wat vir woordeboekgebruikers ter sake is. In die vasstelling van betekeniskenmerke wat in woordeboekverklarings aangebied word, laat leksikograwe hulle grootliks lei deur insigte vanuit ouer benaderings in die linguïstiek. Gevolglik word 'n streng skeiding tussen linguïstiese en ensiklopediese data dikwels steeds gehandhaaf. Na 'n oorsig oor aspekte van woordeboektipologie, woordeboekstrukture en semantiese data in gedrukte woordeboeke word daar in hierdie artikel voorstelle gemaak vir die benutting van nuwe woordeboekstrukture om 'n omvattender weergawe van semantiese data in aanlyn verklarende woordeboeke te bewerkstellig. In die verskillende semantiese subkommentare van 'n artikel word daar voorsiening gemaak vir 'n horisontale ordening van verskillende vakke wat elk 'n datatipe huisves. Meer ruimte word geskep vir semantiese data en ook vir data wat buite die bestek van tradisionele woordeboekinskrywings val. 'n Sterker fokus word ook geplaas op betekenisbetrekkinge en maniere waarop hierdie betekenisleiding in die semantiese kommentaar aangebied kan word. Nuwe tipes geïntegreerde mikrostrukture word voorgestel om die data op gepaste maniere toeganklik te maak vir gebruikers, en voorstelle word gemaak om inligting ook aan woordeboek-eksterne bronne te onttrek. Semantiese data moet so goed en so omvattend as moontlik in verklarende woordeboeke aangebied word.

Slutelwoorde: BETEKENISPARAFRASE, GEÏNTEGREERDE MIKROSTRUKTUUR, PARALLEL-GEÏNTEGREERDE MIKROSTRUKTUUR, SEMANTIESE DATA, SEMANTIESE KOMMENTAAR, SEMANTIESE SUBKOMMENTAAR, SKAGSTRUKTUUR, SOEKTONNEL, SOEKUNIVERSUM, UITGESTELDE GEÏNTEGREERDE MIKROSTRUKTUUR, VAKKE, VEELFASIGE GEÏNTEGREERDE MIKROSTRUKTUUR, VLAKKE

Abstract: A More Comprehensive Presentation of Semantic Data in Online Dictionaries. Meaning is the type of data most frequently needed by dictionary users. Due to space restrictions printed dictionaries can only offer a limited reflection of those aspects of meaning relevant to dictionary users. In determining features of meaning to be used in explanations of meaning, lexicographers are often guided by insights from older approaches in linguistics. Consequently, a strict division is often still maintained between linguistic and encyclopaedic data. Following an overview

* Met waardering opgedra aan Frikkie Lombard.

of some aspects of dictionary typology, dictionary structures and semantic data in printed dictionaries, this paper offers suggestions for the use of new types of dictionary structures to ensure a more comprehensive presentation of semantic data in online monolingual dictionaries. In the different subcomments on semantics in dictionary articles, provision is made for a horizontal ordering of different compartments to accommodate different types of data. More space is allocated to semantic data and also to data that fall beyond the scope of traditional dictionary entries. A stronger focus is placed on semantic relations and ways in which this type of semantic guidance could be accommodated in the comment on semantics. New types of integrated microstructures are suggested to make data accessible for target users, and suggestions are also made for the retrieval of information from dictionary-external sources. Semantic data should be presented as good and as comprehensive as possible in monolingual dictionaries.

Keywords: COMMENT ON SEMANTICS, COMPARTMENTS, DELAYED INTEGRATED MICROSTRUCTURE, INTEGRATED MICROSTRUCTURE, LEVELS, MULTI-PHASED INTEGRATED MICROSTRUCTURE, PARALLEL INTEGRATED MICROSTRUCTURE, PARAPHRASE OF MEANING, SEARCH TUNNEL, SEARCH UNIVERSE, SEMANTIC DATA, SHAFT STRUCTURE, SUBCOMMENT ON SEMANTICS

1. Inleiding

Semantiese data is dié datatipe waarvoor die meeste gebruikers van algemene eentalige en tweetalige woordeboeke hulle tot hierdie naslaanbronne wend. In tweetalige woordeboeke is die soektog veral gerig op die vind van 'n gepaste vertaalekwivalent om die betekenis van 'n brontaalvorm in die doeltaal uit te druk. In eentalige woordeboeke is dit veral 'n verklaring van betekenis waaraan gebruikers 'n behoefte het. Die aanbod van semantiese data is egter nie beperk tot aanduiders van vertaalekwivalente en betekenisparafrases nie. Na gelang van die tipologiese aard van 'n woordeboek, bied woordeboekartikels op semantiese vlak ook leiding ten opsigte van onder meer betekenisbetrekkinge, deursigtige en ondeursigtige gelede woorde, deiktiese waardes van linguïstiese uitdrukkinge, semantiese gebruiksbeperting asook pragmatiese data. Aangesien algemene eentalige en tweetalige woordeboeke ook vir ander datatipes voorsiening moet maak, lei die ruimteprobleme in gedrukte woordeboeke daartoe dat daar selde 'n uitvoerige aanbod van semantiese data kan wees; selfs in omvattende woordeboeke. Na gelang van die woordeboeksoort maak leksikograwe 'n keuse uit die verskillende tipes semantiese data en bied dan slegs aanduiders van hierdie tipes aan. Die omvang van die weergawe van elk van hierdie aanduidertipes word ook deur die woordeboek se tipologiese beperkinge bepaal. Hierdie keuse word normaalweg op grond van die behoeftes van die teikengebruiker gemaak, maar die meeste woordeboektipes kan beswaarlik aan al die behoeftes van al die voornemende gebruikers voldoen. Alhoewel aanduiders van tipes semantiese data dikwels die besetting van soeksones in 'n woordeboekartikel oorheers, bly dit 'n beperkte aanbod wat lei tot 'n beperkte leksikografiese bewerking van die leksikale items wat deur die lemmas verteenwoordig word.

Aanlyn woordeboeke het nie dieselfde ruimteprobleme as gedrukte woordeboeke nie. Alhoewel die aanbiedingsruimte beperk is weens die skermgrootte van die rekenaar, tablet of slimfoon, laat die bergingsruimte groot hoeveelhede data toe. Leksikograwe van aanlyn woordeboeke kan, waar nodig, dus voorsiening maak vir 'n uitvoeriger aanbod van elke datatipe in vergelyking met die aanbod wat in 'n ooreenstemmende gedrukte woordeboek verskyn.

Hierdie artikel is gerig op die vind van moontlike oplossings om die omvangs-probleem van die aanbod van semantiese data in aanlyn woordeboeke te oorkom. Met inagneming van ander datatipes wat ook in woordeboeke aangebied word, is die fokus in hierdie artikel uitsluitlik op semantiese data. Die aard van die woordeboeksoort, die struktuur van die woordeboekartikel en verskillende soeksones in die artikel as soekgebied word onder die loep geneem.

2. Ter agtergrond

Voordat voorstelle vir 'n nuwe aanbod van semantiese data in aanlyn woordeboeke bespreek word, is dit noodsaaklik om kortliks aandag te gee aan enkele aspekte van gedrukte woordeboeke, by name woordeboektipologie, woordeboekstrukture asook enkele aspekte van semantiese data in gedrukte woordeboeke. Hierdie laasgenoemde aspek word voorafgegaan deur enkele verwysings na die semantiek as linguistiese studierigting.

2.1 Aspekte van woordeboektipologie

In die inleidende paragraaf is daar verwys na semantiese data in algemene eentalige en tweetalige woordeboeke. Hierdie verwysing het veral te make met die soort tipologiese indeling wat op gedrukte woordeboeke gerig is. Woordeboektipologie vorm 'n wesenlike deel van leksikografiese besprekings. Scerba (1940) het reeds gewys op die belang van tipologiese verskille tussen woordeboeke. Zgusta (1971: 198, e.v.) bied 'n uitvoerige tipologiese klassifikasie wat as basis dien vir talle latere besprekings van woordeboektipologie.

Die vroeë klassifikasie van woordeboeksoorte was uiteraard op gedrukte woordeboeke gerig en dit geld vandag steeds in 'n hoë mate — ook in talle beskouing van die aanlyn leksikografie. Die tradisionele woordeboekklassifikasie het woordeboeke nie in waterdigte kompartemente geplaas nie, maar wel voorsiening gemaak vir oorvleueling tussen verskillende tipes. Desondanks was daar duidelike kriteria waarvolgens enige woordeboek in 'n bepaalde kategorie geplaas is. Verskillende woordeboeksoorte het verskillende gebruikersgroepe gehad en die aard en aanbod van die verstrekte data is in terme van die woordeboeksoort, die teikengebruikers, en die leksikografiese behoeftes en woordeboekvaardighede van die teikengebruikers bepaal. Algemene eentalige en tweetalige woordeboeke is in die eerste plek as linguistiese instrumente gesien wat linguistiese data aan hulle gebruikers beskikbaar moes stel. Die keuse en aanbod van data is grootliks deur die woordeboeksoort bepaal en dit het beperkte geleentehede

geskep vir verandering en vernuwing.

Wat die hantering van betekenisaspekte in eentalige verklarende woordeboeke betref, het die onderskeid tussen omvattende woordeboeke, standaardwoordeboeke en pedagogiese woordeboeke aan elk van hierdie woordeboeksoorte 'n bepaalde datapakket voorgehou waaraan die leksikograwe hulle moes hou.

2.2 Aspekte van woordeboekstrukture

In die aanbod van semantiese data in woordeboeke speel verskillende strukture 'n rol. Vir die plasing en vind van semantiese data is die dataverspreidingsstruktuur, die toegangstruktuur en die data-onttrekkingstruktuur van die hoogste belang. Vir die doel van hierdie bespreking is dit egter veral die artikelstruktuur en sy onderskeie komponente en elemente wat ter sprake gebring word.

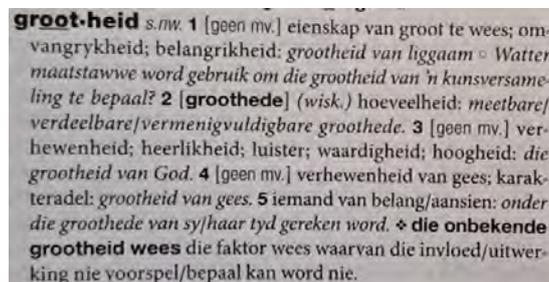
2.2.1 Artikelstruktuur

Algemene eentalige en tweetalige woordeboeke vertoon 'n vaste artikelstruktuur met data wat veral in óf die vorm- óf die semantiese kommentaar gehuisves word. Woordeboekartikels bevat twee tipes elemente, te wete aanduiders en struktuurmerkers (vgl. Wiegand 1989a en Wiegand en Smit 2013). Aanduiders is mikrostrukturele elemente waaraan leksikografiese inligting onttrek kan word. Struktuurmerkers is elemente waaraan geen leksikografiese inligting onttrek kan word nie, maar wat bepaalde datatipes en soeksones merk. Semantiese data word deur middel van aanduiders in woordeboekartikels weergegee, en hierdie datatipe word deur middel van struktuurmerkers van ander datatipes onderskei.

Die semantiese kommentaar bevat verskillende tipes semantiese data soos die aanduiders van die betekenisparafrase, antonieme, sinonieme en, in die geval van tweetalige woordeboeke, vertaalekwivalente. In die geval van lemmas wat 'n polisemiese leksikale item verteenwoordig, bevat die semantiese kommentaar 'n semantiese subkommentaar vir elkeen van die verskillende polisemiese waardes. 'n Semantiese subkommentaar bevat die tersaaklike betekenisparafrase of vertaalekwivalent. Die semantiese kommentaar, sowel as die semantiese subkommentare, bied ook ruimte vir die weergawe van koteksaanduiders. Sulke aanduiders word nie net ter bevrediging van 'n teksproduksiefunksie verstrekk nie, maar ook om die teksbegripfunksie ter wille te wees. Daarom is dit belangrik dat daar 'n duidelike verband tussen 'n betekenisparafrase of vertaalekwivalent en die verstrekte koteksinskrywings moet wees. Vir 'n ondubbelsinnige vergestaltung van hierdie verband word daar voorsiening gemaak vir verskillende tipes mikrostrukture. Dit is veral die geïntegreerde en die niegeïntegreerde mikrostrukture (Wiegand 1989a: 483, 488) wat hier ter sake is en wat die verandering teenwerk wat spruit uit die gebruik van 'n primitiewe mikrostruktuur (vgl. Gouws 2003: 107).

In die geval van 'n geïntegreerde mikrostruktuur tree een of meer koteks-

aanduiders in dieselfde semantiese subkommentaar, 'n integraat (sien Wiegand 1989a: 484 en Wiegand en Smit 2013: 176), as die betekenisparafrase of vertaal-ekwivalent op. Figuur 1, 'n deelartikel uit die HAT, bevat vyf semantiese subkommentare en telkens word die betekenisparafrase-aanduiders aangevul deur een of meer koteksaanduiders wat in dieselfde integraat geïntegreer is. Die nabyheid van koteksaanduiders aan die betekenisparafrase-aanduiders skep 'n verhouding van onmiddellike adressering wat vir die gebruiker 'n kitsonttrekking van inligting moontlik maak.



groot-heid *s.nv.* **1** [geen mv.] eienskap van groot te wees; omvangrykheid; belangrikheid: *grootheid van liggaam* ◦ *Watter maatstawwe word gebruik om die grootheid van 'n kunsversameling te bepaal?* **2** [**grootheide**] (*wisk.*) hoeveelheid: *meetbare/verdeelbare/vermenigvuldigbare grootheide*. **3** [geen mv.] verhewenheid; heerlikheid; luister; waardigheid; hoogheid: *die grootheid van God*. **4** [geen mv.] verhewenheid van gees; karakteradel: *grootheid van gees*. **5** iemand van belang/aansien: *onder die grootheide van sy/haar tyd gereken word*. ♦ **die onbekende grootheid wees** die faktor wees waarvan die invloed/uitwerking nie voorspel/bepaal kan word nie.

Figuur 1: Uit die HAT

In die geval van 'n niegeïntegreerde mikrostruktuur word al die koteksinskrywings wat vir elk van die betekenisonderskeidings verstrek word saam in een teksblok geplaas wat gewoonlik op die teksblokke volg wat die verskillende semantiese subkommentare huisves. Sien in hierdie verband figuur 2, 'n artikel uit die *Prisma Groot woordenboek Afrikaans en Nederlands*:



rond (voorz.), **rond** (voors.) **1** (om iets heen) *om, rond* **2** (in de buurt van, omheen) *rondom, om en by* **3** (met betrekking tot) *rondom* ♦ **1** rond het vat zitten stalen banden *om die vat is daar staal bande*; rond de tafel zitten *om die tafel sit*; er staat een wisselend magnetisch veld rond de stroomdraad *daar is 'n wisselende magnetiese veld om die stroomdraad*; ze renden de kamer rond *hulle het deur die kamer rondgehardloop* **2** het zijn enkele van de vergeten dorpen rond Dokkum *dit is 'n paar van die vergete dorpe rondom Dokkum*; een band van geladen deeltjes rond de aarde *'n band van gelaide deeltjes rondom die aarde*; rond de klok van vijf uur *om en by vyf uur*; de koers schommelde rond de honderd punten *die koers het om honderd punte rond geskommel*; zij is rond de dertig *sy is om en by dertig*; rond Pasen *om en by Paasfees* **3** rond het verschijnsel bestaan verschillende legenden *rondom die verskynsel bestaan daar verskeie legendes*; de geruchten rond zijn benoeming *die gerugte rondom sy benoeming*

Figuur 2: Uit die *Prisma Groot woordenboek Afrikaans en Nederlands*

Die artikel van die lemma *ronde* vertoon drie semantiese subkommentare. Elkeen bevat, in hakies, 'n bondige betekenisverklaring, gevolg deur die tersaaklike vertaalekwivalent(e). Na afloop van die derde semantiese subkommentaar is daar 'n diamantvormige nietipografiese struktuurmerker wat ter wille van kitstoe-gang die artikeldeel merk waar die koteksaanduiders verskyn. Hierna is daar vir elk van die voorafgaande semantiese subkommentare 'n artikelgleuf wat die gepaste koteksaanduiders bevat. Elk van hierdie artikelgleuwe het 'n nommer wat ooreenstem met 'n nommer, as polisemiemerker, van een van die vooraf-gaande semantiese subkommentare. Daardeur word elke koteksgleuf verbind met 'n semantiese subkommentaar wat dieselfde nommer het. Alhoewel elke semantiese subkommentaar nie koteksaanduiders as deel van die betrokke inte-graats bevat nie en daar 'n verhouding van afstandsadressering tussen 'n koteks-aanduiders en die tersaaklike vertaalekwivalent bestaan, maak die eksplisiete nume-ringsafparing dit vir gebruikers moontlik om probleemloos die tersaaklike koteks-aanduiders vir 'n bepaalde vertaalekwivalent te vind.

2.3 Enkele aspekte van die semantiek

Algemene verklarende woordeboeke word veral geraadpleeg vir die onttrek-king van taalkundige inligting. Dit is veral data wat met die betekenis van 'n leksikale item te make het waarna gesoek word en dan veral daardie aandui-ders wat 'n verklaring van die betekenis van 'n woord bied — die aanduiders van betekenisparafrases. As houers van taalkundige data is woordeboeke van die afgelope dekades daarop gerig om 'n suiwer linguistiese verslag van die lek-sikon te bied. Ook wat die weergawe van semantiese data betref, is daar gepoog om nielinguistiese data uit die betekenisparafrase-aanduiders te weer. Woorde-boeke is skerp gekritiseer indien daar 'n vermenging van linguistiese en ensiklo-pediese data in die verklaring van betekenis aangebied is. Vergelyk in hierdie verband die kritiek op betekenisverklaring in vroeër dele van die WAT, byvoor-beeld Combrink (1979) en Odendal (1979).

Volgens Gouws (om te verskyn) poog natuurlike semantiese metataal, 'n teoretiese benadering in die linguistiek, om betekenis in te kort tot 'n beperkte stel semantiese primitiewe. Semantiese primitiewe is die eenvoudigste linguis-tiese konsepte. Hulle kan nie gedefinieer word deur eenvoudiger terme te gebruik nie (vgl. Wierzbicka 1972 en Levisen en Waters 2017). Om betekenis met behulp van semantiese primitiewe te identifiseer, lei tot 'n suiwer linguistiese beteke-nisverklaring wat universele waarde het. Fodor (1977: 155-156) beweer dat in generatiewe grammatika "a lexical item is substituted for a constituent of a phrase marker whose terminal elements are semantic primitives". Met behulp van sulke primitiewe kan die betekenis van die woord *doodmaak* beskryf word as "veroorzaak om nielewend te word". Ook in die strukturele semantiek lei die benadering van komponentanalise daartoe dat 'n woord se betekenis in kom-ponente verdeel word wat slegs ruimte laat vir linguistiese data en geensins vir gebruiks- of ensiklopediese data nie. Kenmerkend hier is die weglating van alle

persoonlike en nielinguistiese kenmerke.

Gouws (om te verskyn) wys daarop dat die kognitiewe semantiek 'n ander benadering volg. Hiervolgens is taal gewortel in die individu se ervaring, insluitend sintuiglike ervarings. 'n Betekenisverklaring wat op hierdie benadering gebaseer is, strek dus verder as semantiese primitiewe en komponente en bevat ook buitelinguistiese kenmerke. Volgens DiMuccio-Failla (2025: 27) se kognitiewe benadering tot betekenis, kan woorde beskou word as verteenwoordigers van ervaringskategorieë. Dit het noodwendig 'n invloed op betekenisverklaring en waar hierdie benadering in woordeboeke gevolg word, sal die betekenisparafrase veel meer insluit as die data wat deur semantiese primitiewe en komponente weergegee word. Klotz en Herbst (2016: 72) het reeds aangedui dat 'n woordeboekdefinisie 'n beskrywing is van die betekeniskenmerke van 'n leksikale item gebaseer op sy gebruik in 'n taalgemeenskap.

Uit die voorafgaande bespreking kan die afleiding gemaak word dat woordeboeke wat slegs op 'n streng linguistiese interpretasie van betekenis gerig is, nie noodwendig die volle omvang van dit wat die taalgemeenskap as betekenis ervaar, weergee nie. Nuwe benaderings in die aanlyn leksikografie maak voorsiening vir aanpassings in hierdie verband.

2.4 Semantiese data in gedrukte woordeboeke

In die hieropvolgende paragrafe word enkele aspekte van semantiese data in gedrukte verklarende woordeboeke bespreek. Dit bied 'n agtergrond vir die bespreking in 'n daaropvolgende afdeling oor semantiese data in aanlyn woordeboeke.

2.4.1 Komponentanalise

Linguistiese teorieë is ten dele ter sake in die leksikografie. Die leksikografie is egter 'n dissipline uit eie reg en nie 'n subdissipline van die linguistiek nie. Daarom speel linguistiese teorieë nie 'n bepaalde rol in besluite oor die opname en aanbieding van data in woordeboeke nie, alhoewel aspekte van sulke teorieë dikwels wel deur leksikograwe oorweeg word. Bepaalde benaderings in die taalwetenskap kan aangepas word vir 'n spesifieke leksikografiese gebruik.

'n Voorbeeld van so 'n aanpassing is te vinde in die linguistiese siening van semantiese primitiewe en komponentontleding. In die linguistiek lei hierdie benaderings tot 'n suiwer taalkundige analise van woordbetekenis. Die individu se ervaring van betekenis asook Wittgenstein (1953) se siening dat die betekenis van 'n woord deur sy gebruik in die taal bepaal word, word nie deur so 'n linguistiese benadering in ag geneem nie. In die leksikografiese hantering van semantiese data kan die benutting van byvoorbeeld betekenis Komponente 'n waardevolle hulp wees, maar 'n toepassing daarvan moet nie die eindpunt van die semantiese analise wees nie. Hier is 'n aanpassing nodig om vir die eiesoortigheid van leksikografiese betekenisverklaring voorsiening te maak. Na aan-

leiding van die werk aan Louw en Nida (1988), dit is die innoverende Grieks–Engelse woordeboek van die Nuwe Testament wat op semantiese velde gebaseer is, stel Louw (1985) 'n benadering bekend wat linguistiese teorie vir die leksikografie aanpas. In die analise van betekenis word semantiese komponente gebruik — Louw werk met drie komponenttipes, te wete gemeenskaplike komponente, diagnostiese komponente en supplementêre komponente. Die gebruik van die eersgenoemde twee tipes lei tot 'n produk vergelykbaar met dié vanuit die linguistiese benadering van komponentanalise. Hierdie analise word dan, waar nodig, aangevul met die gebruik van supplementêre komponente. Hierdie komponenttipe bied data wat nie streng semanties is nie en wat iets van 'n woord se gebruik of sy ensiklopediese waarde weergee. Die insluiting van supplementêre komponente bied 'n leksikografiese werkswyse wat die tradisionele komponentontleding van die strukturele taalkunde verbind met aspekte van die kognitiewe semantiek om 'n gepaste woordeboekbewerking moontlik te maak.

Die gebruik van supplementêre komponente het 'n invloed op die aard en omvang van die betekenisparafrase aangesien bykomende, selfs nielinguistiese, data daarby ingesluit word. Betekenisparafrases in die artikels van gedrukte woordeboeke bied ongelukkig dikwels nie voldoende ruimte vir 'n data-aanbod wat alle gebruikers se kennisbehoefte bevredig nie; veral omdat die betekenisparafrase nie 'n genoegsame hantering van supplementêre komponente toelaat nie. Dié situasie kan ingrypend verbeter word in aanlyn woordeboeke.

2.4.2 Betekenisbetrekkinge

Die leksikale semantiek sluit meer in as die betekenisverklaring van leksikale items. Die leksikon vertoon 'n netwerk van semantiese betrekkinge wat deel is van die studie-objek van die leksikale semantiek. As draers van betekenisdata moet verklarende woordeboeke ook van betekenisbetrekkinge verslag doen. Aan betrekkinge soos antonimie en sinonimie gee verklarende woordeboeke gereeld aandag, maar die spesifieke aard van 'n bepaalde sinonimiese betrekking, byvoorbeeld of dit absolute al dan gedeeltelike sinonieme is wat aangedui word, word selde vermeld. Koteksaanduiders om die voorkoms van die tersaaklike sinonimiese betrekking te illustreer, ontbreek meestal ook.

Naas sinonimie vorm hiponimie deel van die verskillende tipes betekenisbetrekkinge van semantiese insluiting. Waar woordeboeke die genus en differentia-metode gebruik in die betekenisparafrase van 'n woord wat 'n hiponiem is, is daar 'n implisiete aanduiding van 'n hiponimiese verhouding deurdat die superordinaat as gemeenskaplike komponent verstrekkend word. Dit is duidelik waar HAT die betekenis van die woord *kool* verklaar as "enigeen van die groentesoorte en -kultivars van die geslag *Brassica* met dikkerige groen of rooi blare wat styf inengevleg is en 'n harde kop vorm". Hier tree *groentesoort* as superordinaat van *kool* op. Hoofsaaklik weens ruimtebeperkings ontbreek breër semantiese leiding ten opsigte van ander hiponieme van dieselfde superordinaat in die meeste woordeboekartikels. In die betekenisparafrase van 'n woord wat 'n

superordinaat is, word voorbeelde van hiponieme dikwels genoem. Vergelyk HAT se verklaring van *groente* as: "plante/plantdele wat 'n mens rou of gaargemaak (kan) eet, soos kool, blomkool, aartappels, (geel)wortels en ertjies ...". Hier word meer semantiese leiding verstrekk, maar gebruikers kry steeds slegs 'n beperkte deel van die betrokke semantiese veld.

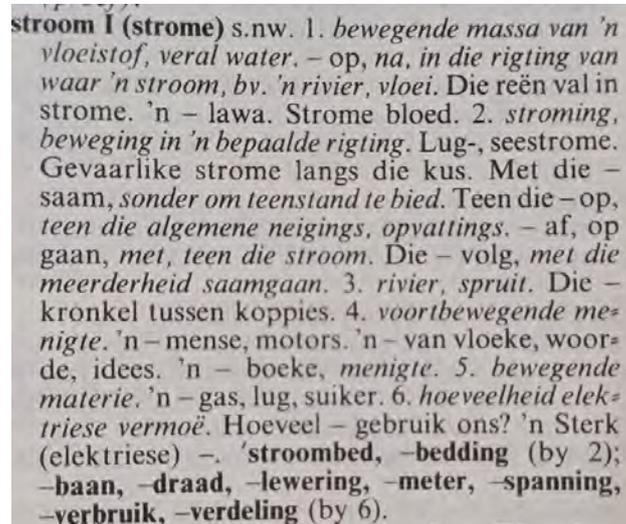
Antonimie is een tipe betekenisteenoorgesteldheid. Woordeboeke toon dikwels antonieme aan of hulle skep 'n soeksone vir die weergawe van betekenisteenoorgesteldes. Vir die woordeboekgebruiker op soek na inligting oor tipes teenoorgesteldes is die tipiese leksikografiese bewerking onvoldoende. Daar word nie genoegsaam inligting verstrekk om die gebruiker te help om tussen byvoorbeeld antonimie, omgekeerdheid en komplementariteit (Lyons 1977: 270-280) te onderskei nie.

2.5 Onverklaarde gelede woorde

Eentalige verklarende woordeboeke het dikwels die gebruik om onverklaarde gelede woorde op te neem. Hierdie woorde kwalifiseer as lemmakandidate, maar weens ruimtebeperkings kry hulle geen volledige bewerking nie. Die implikasie van opname as onverklaarde lemma is dat die gelede woord semanties deursigtig is en dat die betekenis afgelei kan word van die betekenis van die samestellende dele. Deursigtigheid is nie altyd maklik vasstelbaar nie en leksikograwe mag hulle nie op die intuïsie van hulle gebruikers beroep en aanvaar dat dit wat vir die leksikograaf deursigtig is ook vir die gebruiker sal wees nie.

Begripsprobleme duik maklik op, onder meer waar die woordvariant van een of meer stamme van 'n gelede woord polisemies is. Die WAT bied byvoorbeeld agt betekenisonderskeidings en 'n reeks subonderskeidings in die bewerking van die selfstandige naamwoord *kop*. In die lys onverklaarde lemmas verskyn onder meer *kopprofiel* en *kopskub*. Behalwe dat die betekenis van hierdie woorde nie sonder meer deursigtig is nie, het die gebruiker ook geen duidelikheid oor watter een van die betekenisonderskeidings van die lemma *kop* in elke onverklaarde gelede woord ter sake is nie. Vergelyk ook die bespreking van hierdie probleem in Gouws (1989: 77, e.v.).

Sommige woordeboeke bied meer hulp in hierdie verband, byvoorbeeld *Nasionale woordeboek* wat in sy aanbieding van onverklaarde lemmas 'n benadering van neslemmatisering (vgl. Wiegand 1989b: 391) volg, en telkens aantoon watter polisemiese waarde van die woord wat deur die voorafgaande hooflemma verteenwoordig word in die eerste stam van die gelede woord geaktiveer word. Dit is duidelik uit die bewerking in figuur 3. Die artikel van die lemma *stroom* is sesvoudig polisemies. In die lys onverklaarde lemmas word met behulp van 'n deeltrajekpolisemiemerker "(by 2)" aangetoon dat die tweede polisemiese waarde ter sake is by die onverklaarde lemmas *stroombed* en *stroombedding*, terwyl polisemiese waarde ses geaktiveer word in die sublemmas *stroombaan* — *stroomverdeling*.



stroom I (strome) s.nw. 1. *bewegende massa van 'n vloeistof, veral water. – op, na, in die rigting van waar 'n stroom, bv. 'n rivier, vloei. Die reën val in strome. 'n – lawa. Strome bloed.* 2. *stroming, beweging in 'n bepaalde rigting. Lug-, seestrome. Gevaarlike strome langs die kus. Met die – saam, sonder om teenstand te bied. Teen die – op, teen die algemene neigings, opvattinge. – af, op gaan, met, teen die stroom. Die – volg, met die meerderheid saamgaan.* 3. *rivier, spruit. Die – kronkel tussen koppies.* 4. *voortbewegende menigte. 'n – mense, motors. 'n – van vloeke, woorde, idees. 'n – boeke, menigte.* 5. *bewegende materie. 'n – gas, lug, suiker.* 6. *hoeveelheid elektriese vermoë. Hoeveel – gebruik ons? 'n Sterk (elektriese) –. 'stroombed, –bedding (by 2); –baan, –draad, –lewering, –meter, –spanning, –verbruik, –verdeling (by 6).*

Figuur 3: Uit *Nasionale woordeboek*

'n Ondubbelsinnige vasstelling van die spesifieke betekeniswaarde van onverklaarde lemmas is nie probleemloos nie. Hier moet leksikograwe aandag gee aan Gove (1966: 183) wat aanvoer dat die *self-* in *self-explanatory* na die interpreterder van die woord moet verwys en nie na die woord as sodanig nie — vergelyk ook Gouws (1989: 78).

In talle verklarende woordeboeke skiet die implisiete betekenisleiding by die weergawe van onverklaarde gelede woorde te kort. Teikengebruikers van 'n betrokke woordeboek kan nie sonder meer die betekenisinligting aan die aangebode data onttrek nie. Dit striem suksesvolle semantiese leiding in die artikels van verklarende woordeboeke.

3. Aanlyn woordeboeke

3.1 Woordeboektipologie

Tarp (2009: 24) onderskei tussen 'n kontemplatiewe en 'n transformatiewe benadering in die leksikografie. Eersgenoemde bring mee dat leksikograwe op dieselfde manier voortwerk aan hulle woordeboeke as hulle voorgangers, terwyl die laasgenoemde benadering daartoe lei dat leksikograwe vernuwend optree in die beplanning en samestelling van woordeboeke. Ook in die oorgang van gedrukte na aanlyn woordeboeke is daar dikwels 'n kontemplatiewe benadering met dieselfde werkswyse wat gehandhaaf word. Dit geld ook die tipologiese klassifikasie, want die indeling wat vir gedrukte woordeboeke geld, kan sonder meer of met slegs geringe aanpassings ook in die aanlyn omgewing

gebruik word. Die aanlyn omgewing het egter 'n transformatiewe potensiaal en bied moontlikhede wat tot wesenlike aanpassings in die bestaande woordeboektipologie kan lei. Dit het implikasies vir baie aspekte van die leksikografiepraktyk, onder meer vir die data-aanbod en die dataverspreidingstruktuur.

In die aanlyn leksikografie speel toegang tot data 'n deurslaggewende rol. Belangriker as die beskikbaarheid van 'n bepaalde woordeboeksoort is die beskikbaarheid van die gepaste data — ongeag of dit in 'n woordeboek of in 'n databasis gevind word. Toegang tot data kan ook geïndividualiseer word sodat elke gebruiker 'n unieke toegang tot die spesifieke data kry waaraan hulle 'n behoefte het. In hierdie verband bied Gouws (2006) en Bergenholtz en Bergenholtz (2017) aanduidings van meerdere woordeboeke wat aan een databasis onttrek kan word, terwyl Bergenholtz, Bothma en Gouws (2011) wys op toegang tot vooraf geselekteerde datatipes. Die beginsels wat in hierdie publikasies ter sprake gebring is, kan in die hedendaagse aanlyn leksikografie met veel meer verfyning en presiesheid toegepas word — ook wat die beskikbaarstelling van semantiese data betref. Dit kan lei tot 'n vernuwende aanpassing in die tipologiese klassifikasie van leksikografiese bronne.

Aan 'n noukeurig beplande en gemerkte databasis kan verskillende woordeboeke onttrek word. Die data in die databasis van so 'n "moederwoordeboek" (vergelyk Gouws 2006), kan gemerk word vir spesifieke woordeboeksoorte, byvoorbeeld 'n eentalige verklarende skoolwoordeboek, aanleerderwoordeboek, standaardwoordeboek of omvattende woordeboek. Na gelang van die gebruikersprofiel, soos aangedui deur 'n gebruiker wanneer die databasis benut word, kan verskillende bewerkings van dieselfde lemma gebied word. Vir teksbegripdoeleindes kan verskillende aanduiders van die betekenisparafrase aan die databasis onttrek word en vir teksproduksiedoeleindes kan verskillende koteksaanduiders verskyn. Tradisionele woordeboektipologie is hier nie so belangrik nie, maar wel die behoeftes van die gebruiker — en dit word in die gebruikersprofiel gespesifiseer. Waar 'n gebruiker se profiel hom/haar as 'n aanleerder van 'n taal spesifiseer, sal die tersaaklike aanduiders onttrek word. Aan dieselfde databasis onttrek die gebruiker wat as moedertaalspreker met 'n behoefte aan omvattende betekenisleiding gespesifiseer is, ander tersaaklike inligting.

Die databasis kan ook so opgestel word dat woordeboeke wat slegs enkele datatipes bevat daaraan onttrek kan word. Die databasis bevat al die leksikografiese data vir 'n woordeboek wat die volle spektrum van datatipes dek, maar 'n gebruiker wat slegs 'n behoefte aan, byvoorbeeld, uitspraakhulp het, onttrek slegs aanduiders wat die uitspraak van 'n betrokke woord weergee. Op dieselfde manier kan 'n woordeboek onttrek word wat slegs semantiese data bevat. Telkens kan die leksikograaf aandui watter tipes aanduiders ter sake is. In die geval van 'n uitspraakwoordeboek kan dit byvoorbeeld 'n fonetiese transkripsie van die woord wees of dit kan primêre en sekondêre klem, asook sillabeverdelings aantoon. In die geval van 'n betekeniswoordeboek is verskeie aanduidertipes moontlik, byvoorbeeld 'n betekenisparafrase, antonieme en sinonieme.

3.3 Woordeboekstrukture

Aanlyn woordeboeke vertoon 'n verskeidenheid strukture wat ooreenstem met asook verskil van strukture in gedrukte woordeboeke. Sommige strukture, byvoorbeeld die makrostruktuur, word met bepaalde aanpassings gehandhaaf. Ander strukture, byvoorbeeld die raamstruktuur, val weg in aanlyn woordeboeke, terwyl nuwe strukture, byvoorbeeld die skermkootstruktuur, na vore tree. Dit is veral die artikelstruktuur en strukture wat in en vanuit die artikel ingespan word wat vernuwingsmoontlikhede in die data-aanbod, ook dié van semantiese data, tot voordeel kan strek.

3.3.1 Die artikelstruktuur in aanlyn woordeboeke

In die hieropvolgende bespreking is die fokus uitsluitlik op aspekte van die semantiese kommentaar van die artikels van aanlyn woordeboeke. In die semantiese kommentaar vertoon aanlyn woordeboeke dikwels dieselfde statiese struktuur wat in gedrukte woordeboeke voorkom. Vergelyk in hierdie verband figuur 4, die artikel van die lemma *taksonomie* in die aanlyn WAT waarvan die statiese artikelstruktuur baie ooreenstem met dié van artikels in die gedrukte dele van die WAT:

taksonomie (tak-so-no-mie) s.nw. (met mv. in bet. 2 b en 3 b) (†Gr. taxis, "rangskikking" en nomos, "norm, reël")

1 Wetenskap of leer van klassifikasie, asook die beginsels en wette wat binne sodanige klassifikasie geld.

Vgl. [SISTEMATIEK 1](#).

2 a Afdeling v.d. biologie wat gerig is op die benaming, beskrywing en klassifikasie van lewende organismes volgens verwantskappe en verskille: *Reeds in Linnæus se tyd is sowel kwantitatiewe as kwalitatiewe data in die taksonomie gebruik* (J. Heyns: Taksonomie, 1973, 15). *Die komitee doen ... 'n voortdurende studie oor die wêreld se wild, hou by met die menings wat taksonomie betref, en hersien die kategorieë van trofeediëre* (J. Lötter, vert.: Jag, 2002, 146).

Vgl. [SISTEMATIEK 2](#).

b mv. *taksonomieë*. Voorbeeld van 'n klassifikasie binne die taksonomie ([TAKSONOMIE 2 a](#)): *Die grootste deel van my lewe was ek besig met die taksonomie van ons insekte* (Volksbl., 23 Mei 2017, E). *Bespreking van die taksonomie en beskrywings van die anatomie en uiterlike voorkoms van die bontebok is te vinde* (Koedoe, 1976, 67). *Goeie wetenskaplike versamelings en taksonomieë het .. voortgespruit uit studies van blaarvloë en blaarspringers, parasietwespe, dopluise en sprinkane* (Panorama, Des. 1987, 39).

Figuur 4: Uit die aanlyn WAT

Hierdie artikel vertoon 'n mikrostrukturele ordeningstruktuur met aanduiders in die semantiese kommentaar wat vertikaal georden is. 'n Vertikale hiërargiese ordening kom algemeen in woordeboekartikels voor en in gedrukte woordeboeke bied dit gemaklike interne soekroetes. 'n Vertikale ordening lei tot 'n skagstruktuur met die gebruiker wat in die interne toegangstruktuur 'n soekroete van bo na onder volg.

Aanlyn woordeboeke vertoon soms ook 'n horisontale ordening in die arti-

kelstruktuur, soos te sien in figuur 5 uit die ANW. Die linkerkantste kolom in die skermkoot vertoon die makrostrukturele deeltrajek waarbinne die lemma *hondenfluitje* pas. Die volgende kolom is die begin van die artikel en bevat 'n spyskaart van datatipes en ander inskrywings waartoe die gebruiker in en vanuit die artikel toegang het. Die daaropvolgende kolom bied 'n gedeeltelike bewerking van die eerste polisemiese waarde van *hondenfluitje*, terwyl die regterkantste kolom bepaalde grammatiese aanduiders bevat.

The screenshot shows the ANW website interface for the word 'hondenfluitje'. The left sidebar contains a vertical list of letters from A to Z, with 'H' highlighted. The main content area features the title 'hondenfluitje' and a definition: 'hondenfluitje 1.0 fluitje met een heel hoog geluid om honden mee te roepen of ze andere commando's te geven'. Below the definition is a 'Semagram' section and a 'Doel of bestemming' section. The right sidebar contains grammatical information, including 'Woordsoort' (zelfstandig naamwoord), 'Naamtype' (soortnaam), 'Geslacht' (onzijdig), 'Lidwoord' (het), and 'Betekenisklasse' (zaaknaam). There is also a 'Spelling en flexie' section with a table showing the singular and plural forms of the word.

Vorm	Afleiding
Enkelvoud	hondenfluitje hon.den.fluit.je
Meervoud	hondenfluitjes hon.den.fluit.jes

Figuur 5: Uit die ANW

Dit is 'n dinamiese artikelstruktuur — 'n klik op 'n spyskaartitem vul die middel- en/of regterkantste blok met die tersaaklike nuwe data. 'n Klik op byvoorbeeld die spyskaartinskrywing *Voorbeelden*, lei tot figuur 6 waar die regterkantste blok van figuur 5 verdwyn en die oorspronklike middelblok, nou die regterkantste blok, deur die bondige betekenisfrase en meegaande voorbeeldaanduiders beset word.

Die voorbeeldaanduiders wat hier verstrekk word, is ter illustrasie van die spesifieke betekenisonderskeiding van die woord wat deur die lemma verteenwoordig word. Dit word in dieselfde integraat as die betekenisparafraase-aanduiders verstrekk en gesamentlik vertoon om as voorbeeld van 'n geïntegreerde mikrostruktuur te dien. Anders as in die geval van figuur 4, die voorbeeld uit die WAT, is die voorbeeldaanduiders nie deel van die eerste skermkoot waarin betekenisparafraase 1 vertoon word nie. Die voorbeeldaanduiders word eers na 'n verdere klik sigbaar in die integraat. Hier is dus sprake van 'n uitgestelde geïntegreerde mikrostruktuur. Die sigbaarmaking van die geïntegreerde mikrostruktuur word uitgestel tot na die klik van 'n data-identifiseerder.

The screenshot shows the ANW website interface. On the left is a navigation menu with letters A through Z. The main content area is titled 'hondenfluitje'. It includes sections for 'Per betekenis:' with sub-sections 1.0 and 2.0, 'Per onderwerp:' with various linguistic categories, 'Voorbeelden', and 'Zoek 'hondenfluitje' ook in:' with links to other resources. The right side of the page contains 'hondenfluitje 1.0' with a definition and 'Algemene voorbeelden' with several example sentences and citations.

Figuur 6: Uit die ANW

Die horisontale plasing van aanduiders en teksblokke kan ook tot verdere tipes geïntegreerde mikrostrukture lei. Vergelyk in hierdie verband figuur 7–10, dele van die Duitse woordeboek *lexico* se hantering van die lemma *Pferd*:

The screenshot shows the 'lexico' dictionary entry for 'Pferd'. It features a yellow square icon. The entry is organized into sections: 'Orthografie' (Normgerechte Schreibung: Pferd, Worttrennung: Dieses Wort ist nicht trennbar.), 'Herkunft und Wandel' (Etymologische Angaben: anzeigen >, Wandel 1700 bis 1945: -, Wandel seit 1945: -), and 'Wortbildungsprodukte' (automatisch ermittelt) weiter >. Below these are 'Lesartenbezogene Angaben' (Lesartenübergreifende Angaben) with three sub-entries: 'Reittier' (weiter >), 'Turngerät' (weiter >), and 'Schachfigur' (weiter >), each with a brief description of the term's usage.

Figuur 7: Uit *lexico*

Die semantiese kommentaar van hierdie artikel bevat verskeie subkommentare as integrale waarin die verskillende polisemiese waardes van die woord *Pferd* bewerk word. Deur 'n vertikale skagstruktuur in die veelvlakkige semantiese kommentaar beweeg 'n gebruiker na die relevante semantiese subkommentaar. Figuur 8 is die produk van 'n klik op die merker *weiter* (= verder) van 'n subkommentaar, hier dié van die betekeniswaarde *Reittier* (= rydier), as beperkte artikel (vergelyk Gouws en Bothma 2024: 358):

elexiko

Pferd

Lesart: 'Reittier'

zur Übersichtsseite Lesarten im Überblick

Bedeutungs-erläuterung Kollo-kationen Konstruk-tionen Sinnverwandte Wörter Gebrauchs-besonderheiten Grammatik

Erläuterung der Bedeutung / Funktion

Mit **Pferd** bezeichnet man ein großes Säugetier mit langen Beinen, das vom Menschen bevorzugt als Reittier und gelegentlich auch als Zug- und Lasttier genutzt wird.

Belege anzeigen » Illustrationen anzeigen »

Wortklasse: Individuativum

Figuur 8: Uit *elexico*

Die betekenisparafrese is die enigste aanduider in hierdie beperkte artikel, maar daar is ook die data-identifiseerders *Belege* (= sitate) en *Illustrationen* (= illustrasies) wat elk deur die merker *anzeigen* (= wys) gevolg word. 'n Klik op die data-identifiseerder *Belege* se *anzeigen* lei tot figuur 9, 'n beperkte artikel wat naas die betekenisparafrese-aanduider ook die voorbeeldaanduiders vertoon, waarvan slegs enkeles in hierdie afdruk van 'n deel van die skermkoot inpas. Ook hier word 'n uitgestelde geïntegreerde mikrostruktuur benut:

elexiko

Pferd

Lesart: 'Reittier'

zur Übersichtsseite Lesarten im Überblick

Bedeutungserläuterung Kollokationen Konstruktionen Sinnverwandte Wörter Gebrauchsbesonderheiten Grammatik

Erläuterung der Bedeutung / Funktion

Mit **Pferd** bezeichnet man ein großes Säugetier mit langen Beinen, das vom Menschen bevorzugt als Reittier und gelegentlich auch als Zug- und Lasttier genutzt wird.

Belege verbergen × Illustrationen anzeigen >

Sie sind die Stars der Spanischen Reitschule in Wien und untrennbar mit der Geschichte Österreichs verbunden - die edlen Lipizzaner. Jetzt steht fest, die **Pferde** bleiben "rotweißrot". Die Würfel im kuriosen Streit um die Lipizzaner sind gefallen: Die weißen **Pferde** - weltweit ein renommiertes Aushängeschild für die hohe Kunst der Dressur - bleiben Österreicher! Bekanntlich hatte Italien die Zucht der Rasse für sich beansprucht. Landwirtschaftsminister Molterer konnte jetzt seinen Kollegen in Rom überzeugen, auf jegliche Forderungen zu verzichten. (Neue Kronen-Zeitung, 16.01.1999, S. 10, Lipizzaner bleiben Österreicher!)

«Seit der Abschaffung der Kavallerie hat ein kompletter Wandel stattgefunden», stellt Reitclub-Präsident Kurt Hättenschwiler fest. Hatte der Reitclub Flawil bei der

Figuur 9: Uit *elexico*

Gouws en Bothma (2024: 360) toon aan hoe die verskillende vlakke wat elk een semantiese subkommentaar bevat verskillende vakke het, geïdentifiseer deur horisontaal-geordende data-identifiseerders, wat per soektonnel bereik kan word (Gouws en Bothma 2024: 362). 'n Klik op enigeen van die data-identifiseerders wat die vakke identifiseer, lei tot 'n beperkte artikel wat die lemma as gidselement bevat asook die eenwoordweergawe van die betrokke betekenisonderskeiding — hier *Reittier*. Die data in elkeen van hierdie vakke is gerig op die gebruik van die woord in die betrokke betekeniswaarde van die tersaaklike semantiese subkommentaar. In hierdie semantiese subkommentaar is dit nie net die voorbeeldaanduiders wat geïntegreer word nie, maar ook byvoorbeeld die kollokasies, konstruksies en verwante woorde. Figuur 10 wys hoe 'n klik op

die data-identifiseerder *Konstruktionen* tot 'n vak lei wat verskeie konstruksies vertoon:

The screenshot shows the 'Pferd' entry in the 'elexiko' dictionary. The word 'Pferd' is displayed in a large font, with the variant 'Lesart: 'Reittier'' below it. A navigation bar at the top includes a link to the overview page and a menu for 'Lesarten im Überblick'. Below this is a horizontal menu with tabs for 'Bedeutungs-erläuterung', 'Kollo-kationen', 'Konstruk-tionen', 'Sinnverwandte Wörter', 'Gebrauchs-besonderheiten', and 'Grammatik'. The 'Konstruk-tionen' tab is active, showing a section titled 'Konstruktionen: Typische Verwendungen'. This section is divided into three sub-sections, each with a 'verbergen x' link: 'Verwendungen mit Attribut' (examples: 'das Pferd als Arbeitstier', 'Pferde im Galopp'), 'Verwendungen in Verbalphrasen und Sätzen' (examples: 'auf ein Pferd setzen', 'vom Pferd fallen', and several sentence templates with placeholders like '[Personenname]' and '[zahl]'), and 'Verwendungen als Attribut' (examples: 'mit Pferd und Wagen', 'alles rund ums Pferd', 'zu Fuß oder mit dem Pferd', 'eine von Pferden gezogene Kutsche', 'von Pferden gezogene Wagen', 'Zucht- und Halteprämien für Pferde', 'Beziehung zwischen Mensch und Pferd', 'Harmonie zwischen Pferd und Reiter').

Figuur 10: Uit *elexiko*

Verskillende tipes semantiese data word in die vakke van 'n enkele vlak aangebied en die dinamiese aard van die artikel maak dit moontlik vir 'n gebruiker om slegs dit te sien waarna hulle op soek is sonder om deur al die aanduiders hoef te worstel.

4. 'n Omvattender aanbod van semantiese data in Afrikaanse aanlyn woordeboeke

Die voorafgaande voorbeelde uit *ellexiko* wys hoe 'n vernuwend gebruik van strukture in die woordeboekartikel tot 'n omvattender data-aanbod kan lei sonder om gebruikers aan data-oormoed bloot te stel. Afrikaanse verklarende woordeboeke kan op 'n soortgelyke manier verskillende vlakke en vakke in woordeboekartikels benut vir die aanbieding van semantiese data. In die hieropvolgende bespreking is die fokus op die aanbod van semantiese data in verklarende aanlyn woordeboeke. Tipologiese subklassifikasie van verklarende woordeboeke word nie hier in ag geneem nie, maar daar word van die standpunt uitgegaan dat leksikograwe oor die aard en omvang van semantiese data besluit na gelang van die tipologiese aard van 'n beplande woordeboek. Die voorstelle wat vervolgens gemaak word, sou ook op 'n enkele moederwoordeboek of databasis gerig kon wees waaraan verskillende woordeboeke onttrek kan word. Dit word egter nie hier bespreek nie.

4.1 Vlakke en vakke

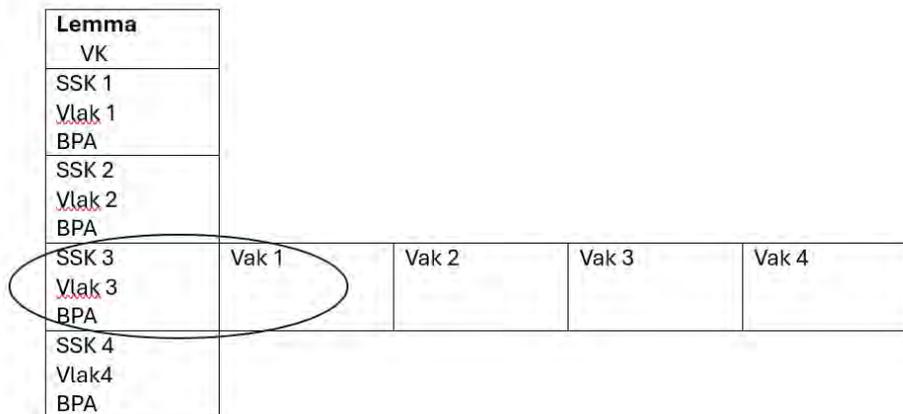
'n Omvattende aanbod van semantiese data vereis 'n artikelstruktuur waarin data-oormoed vermy word, maar waarin vindbare data verstrekkend word waaraan inligting op 'n ondubbelsinnige manier onttrek kan word. Die uitleg van die woordeboekartikel en die meegaande gebruik van bepaalde strukture speel hier 'n beduidende rol. Müller-Spitzer en Koplenig (2014) gee aandag aan verskillende artikeluitlegmoontlikhede. Hulle gebruikersnavorsing het getoon dat gebruikers voorkeur gee aan die tipe uitleg soos in *ellexiko* waar versteekte data deur 'n klik op 'n data-identifiseerder gevind kan word (Müller-Spitzer en Koplenig 2014: 195). Gouws en Bothma (2024: 359-360) verwys in hierdie uitlegtipe na die benutting van vlakke en vakke om vertikale en horisontale soekposisies te skep. So 'n uitleg kan van groot waarde wees in die benutting van 'n dinamiese artikelstruktuur waar alle data nie gelyktydig gesien word nie en waar die gebruiker kan kies wanneer watter aanduiders vertoon moet word. Hier speel sowel 'n skag- as 'n tonnelstruktuur 'n wesenlike rol om die gepaste semantiese subkommentaar en die verlangde datatipe te vind.

Gegee die potensiaal van 'n dinamiese artikelstruktuur en die voordele van 'n vertikale ordening van vlakke wat verskillende semantiese subkommentare huisves en die horisontale ordening van vakke wat verskillende tipes aanduiders bevat wat tot een semantiese subkommentaar behoort, kan leksikograwe aanpassings in die dataverspreidingstruktuur aanbring wat 'n veel uitvoeriger aanbod van semantiese data tot gevolg kan hê. Voorstelle in dié verband word in die hieropvolgende bespreking gemaak.

4.2 'n Veelfasige geïntegreerde mikrostruktuur

'n Basiswoordeboekartikel bestaan uit twee kommentare, naamlik 'n vormkom-

Die semantiese subkommentaar van 'n artikel met 'n veelfasige geïntegreerde mikrostruktuur kan, naas die vertikaalgeordende soekgleuf (dit is die gidsvak waarin die betrokke polisemiese waarde se betekenisparafrase aangedui word), 'n minimum van twee en 'n onbeperkte maksimum vakke bevat met data wat binne die bestek van die betekenisparafrase val. Die geïntegreerde aard van die mikrostruktuur word nie deur die gelyktydige integrasie van die inhoud van al die vakke gesamentlik bepaal nie, maar deur die integrasie, in verskillende fases, van die inhoud van 'n individuele vak en die gidsvak. 'n Klik op die data-identifiseerder van, byvoorbeeld, vak 1, skep 'n geïntegreerde mikrostruktuur, ook 'n uitgestelde geïntegreerde mikrostruktuur soos in die ANW (figuur 6), met vak 1 wat saam met die gidsvak hierdie geïntegreerde mikrostruktuur vorm. Telkens skep die klik op 'n data-identifiseerder 'n onmiddellike adressering tussen die data van daardie vak en die optrede van die leksikale item wat deur die lemma verteenwoordig word in die polisemiese waarde soos deur die betekenisparafrase aangetoon. Figuur 12 wys hoe so 'n integrasie van die gidsvak en 'n horisontaal-geordende vak 'n parallel-geïntegreerde mikrostruktuur vorm.



Figuur 12: 'n Parallel-geïntegreerde mikrostruktuur

4.2.1 Besetting van die vakke

'n Vernuwende aspek van verklarende aanlyn woordeboeke is die dataverspreiding en -aanbod in die verskillende vakke van 'n artikelvlak. Elke vertikaalgeordende vlak wat die bewerking van 'n afsonderlike semantiese subkommentaar huisves, begin met 'n gidsvak wat gevolg word deur horisontaalgeordende vakke wat elk 'n parallel-geïntegreerde mikrostruktuur vertoon. Die gidsvak bevat tipieserwys 'n herhaling van die lemmateken met, indien nodig, een of meer leksikografiese etikette om bepaalde afwykings te merk van die optrede van die woord wat deur die lemma verteenwoordig word in die spesifieke poli-

semiese waarde. Daarnaas word die tersaaklike betekenisparafrase-aanduider ook verstrek. Daar is 'n onmiddellik lemmatiese adressering tussen die betekenisparafrase-aanduider en die lemma wat as gidselement van die gidsvak optree. So 'n lemmatiese adressering geld ook vanuit elke afsonderlike vak wat parallel aan die gidsvak aangebied word.

Die leksikograaf het die vryheid om te bepaal in watter volgorde die data in die verskillende vakke verstrek moet word. Dit moet nie 'n arbitrêre besluit wees nie, maar moet tematiese samehang asook die behoeftes van die aangewese teikengebruiker in ag neem. Deur elke vak van 'n gepaste data-identifiseerder te voorsien, vergelyk figuur 13, kry die gebruiker 'n oorsig oor die datatipes wat in die betrokke semantiese subkommentaar aangebied word. Let wel, slegs enkele datatipes word in hierdie figuur vertoon. Veel meer tipes kan probleemloos in 'n veelfasige geïntegreerde mikrostruktuur geakkommodeer word.

SSK 3	Vak 1	Vak 2	Vak 3	Vak 4	Vak 5
Lemma BPA	Betekenis +	Koteks 1 Voorbeeld sinne	Koteks 2 Kollokasies Voorbeeldsinne	Sinonieme 1. Absoluut Voorbeeldsinne 2. Gedeeltelik Voorbeeldsinne	Teenoorgesteldes Antonieme Omgekeertes Komplementêre vorme

Figuur 13: Besetting van vakke

4.2.1.1 Betekenis +

Hier word die eerste vak beset deur data ter aanvulling van die betekenisparafrase. Soos vroeër in hierdie artikel genoem, is taal gewortel in die individu se ervaring en die individu se interpretasie van 'n woord se betekenis is gebaseer op hierdie ervarings. Van hierdie ervaringsdata val buite die eng linguïstiese benadering van betekenisverklaring waarvan die hoofbetekenisparafrase blyke gee. In 'n vak wat onmiddellik op die gidsvak volg, kan 'n heenkome gebied word aan bykomende, selfs ook buitelinguïstiese, data wat die leksikograaf as belangrik beskou ter ondersteuning en ter aanvulling van die data in die betekenisparafrase. Op hierdie manier word voorsiening gemaak, onder meer deur die benutting van supplementêre kenmerke, vir die bevrediging van 'n breër kognitiewe funksie van die betrokke woordeboek. 'n Parallel-geïntegreerde mikrostruktuur plaas hierdie betekenis +-data in dieselfde integraat as die data in die gidsvak.

4.2.1.2 Koteksleiding

4.2.1.2.1 Voorbeeldsinne

'n Tipiese kenmerk van 'n geïntegreerde mikrostruktuur is die plasing van koteksaanduiders in dieselfde integraat as die betekenisparafrase-aanduider. 'n Klik op die data-identifiseerder van vak 2 skep 'n parallel-geïntegreerde mikrostruktuur, as onderdeel van die veelfasige geïntegreerde mikrostruktuur, wat koteksaanduiders met die betekenisparafrase verbind. Die woordeboekgebruiker wat slegs 'n teksbegripbehoefte het se woordeboekraadpleging hoef nie verder as 'n kyk na die gidsvak te gaan nie. Waar 'n teksproduksiefunksie bevredig moet word, lei 'n klik op die koteksmerker die gebruiker na die ondersteunende voorbeeldsinne — wat poëme en/of sitate kan wees.

Woordeboeke verstrek voorbeeldsinne ter wille van teksproduksiebehoefte, maar voorbeeldsinne ondersteun ook beter teksbegrip. Vir leksikografe is dit nie altyd ewe maklik om te weet of 'n bepaalde optrede van 'n woord 'n eie betekenisonderskeiding verteenwoordig al dan 'n bepaalde gebruik van die woord in 'n bestaande betekenisonderskeiding nie. Behoort die optrede van die woord *kort* in 'n *kort referaat* en 'n *kort rok* tot verskillende betekenisonderskeidings of is dit verskillende gebruike van dieselfde betekenisonderskeiding, te wete "van beperkte afmeting"? Vergelyk ook Gouws (1987). In 'n vak wat voorbeeldsinne bevat, kan 'n groter verskeidenheid voorbeelde verstrek word ter illustrasie van die bepaalde betekenisonderskeiding, maar ook van verskillende gebruike van 'n woord in dieselfde betekenisonderskeiding. Dit verteenwoordig 'n uitbreiding van koteksaanduiders en bied toegevoegde waarde aan die aanlyn woordeboek.

4.2.1.2.2 Kollokasies

Ter wille van omvattender koteksleiding kan dit nuttig wees om tussen voorbeeldsinne en kollokasies te onderskei. Dit word soms ook in gedrukte woordeboeke gedoen, maar dikwels met beperkter aandag aan kollokasies. Daarom bied vak 3 verdere koteksleiding, te wete kollokasie-aanduiders. Die verstrekking van kollokasies kan belangrik wees vir teksproduksiedoeleindes. Kollokasies is nie leksikale items in eie reg nie. Desondanks kan woordeboeke hoër status aan hulle gee as net dié van koteksaanduiders. Naas die gee van kollokasies as koteksaanduiders gerig op die lemma kan verdere koteksaanduiders verskaf word waarin nie net die woord optree wat deur die lemma verteenwoordig word nie, maar die kollokasie as sodanig. Kollokasies word dan verhef tot sekondêre bewerkingsseenhede wat as adres van 'n eie koteksaanduider optree. So 'n koteksaanduider is deel van 'n tweeledige adresseringsverhouding, naamlik 'n lemmatiese adressering asook 'n nielemmatiese adressering met die kollokasie wat die adres is. Die tweede koteksvak in figuur 13 (vak 3) kan kollokasies bevat met die lemma as een van die komponente — hetsy as basis hetsy as kollokator. Naas hierdie kollokasies kan die vak ook voorbeeldmateriaal bevat as koteksaanduiders van die betrokke kollokasies.

'n Vak wat kollokasies bevat asook voorbeeldmateriaal met die kollokasie as adres verteenwoordig 'n verdere tipe geïntegreerde mikrostruktuur. Die onmiddellike adressering van koteksaanduiders en kollokasie skep 'n eie integraat en bied 'n vak-interne geïntegreerde mikrostruktuur. Die geïntegreerde mikrostruktuur wat in die semantiese subkommentaar geld, is die primêre geïntegreerde mikrostruktuur, terwyl die geïntegreerde mikrostruktuur wat slegs binne 'n vak geld as 'n sekondêre geïntegreerde mikrostruktuur beskou kan word.

4.2.1.3 Betekenisbetrekkinge

Vak 4 en 5 is gemerk vir bepaalde tipes betekenisbetrekkinge, te wete sinonimie en semantiese teenoorgesteldes. Die leksikon bevat 'n netwerk betekenisbetrekkinge en in verdere vakke sou daar ook aan ander tipes betekenisbetrekkinge aandag gegee kon word, byvoorbeeld hiponimie. Vergelyk ook die bespreking in 2.4.2.

4.2.1.3.1 Sinonieme

Aanduiders van sinonieme is woordeboekinskrywings waarby gebruikers baie kan baat. In woordeboeke word sinonieme op verskillende maniere en in verskillende artikelgleuwe verstrek, telkens as inskrywings met 'n lemmatiese adressering. Sinoniemaanduiders kry selde in woordeboeke 'n eie bewerking om die aard van die betrokke sinonimiese betrekking duidelik te maak. Woordeboeke onderskei nie konsekwent tussen absolute en gedeeltelike sinonimie nie en dit kan striemend wees vir die kommunikatiewe sukses van die woordeboekgebruiker.

'n Omvattender aanbod van semantiese data behoort ook tot 'n omvattender verslag van betekenisbetrekkinge te lei. In vak 4 sou daar dus tussen absolute en gedeeltelike sinonieme onderskei kan word en gepaardgaande voorbeeldmateriaal sou die korrekte gebruik kon illustreer. As 'n voorbeeld: in die bewerking van die woord *besope* behoort *dronk* nie sonder verdere inskrywings as sinoniem verstrek te word nie. *Besope* het 'n verhouding van gedeeltelike sinonimie met *dronk* en die aanduiders van *dronk* as sinoniem moet vergesel wees van 'n merker van die spesifieke polisemiese waarde van *dronk* wat in 'n sinonimiese betekenisbetrekking tot *besope* staan. Sonder hierdie aanvullende inskrywing sou die gebruik van *dronk* in die betekenisonderskeiding "lighoofdig" ten onregte ook as sinonimies tot *besope* gesien kon word. Naas 'n aanduiding van die tipe sinonimiese verhouding asook van spesifieke betekenisonderskeidings van woorde wat by so 'n verhouding betrokke is, kan voorbeeldmateriaal daartoe bydra om die sinonimiese verhouding te ekspliseer en die gebruiker van beter semantiese leiding te voorsien. Waar sinoniemaanduiders deur koteksaanduiders ondersteun word, vertoon die woordeboek nogmaals 'n vak-intern geïntegreerde mikrostruktuur.

4.2.1.3.2 Betekenisteenoorgesteldes

Binne die breë veld van semantiese opposisie kan verskillende tipes betrekkinge van teenoorgesteldheid onderskei word, byvoorbeeld antonimie, omgekeerdheid en komplementariteit. 'n Woordeboek wat as taalkundenaslaanbron dien, moet 'n presiese hantering van hierdie betrekkinge bied en moet dit vermy om 'n term soos *antoniem* oorkoepelend te gebruik vir die verskillende tipes teenoorgesteldheid.

Vak 5 kan data bevat om verskillende tipes betekenisteenoorgesteldheid voor te stel en gebruikers moet op 'n ondubbelsinnige manier tussen die tipes kan onderskei. 'n Benoeming van die onderskeie tipes betekenisteenoorgesteldes tesame met die gebruik van relevante kotteksaanduiders sal ook 'n waardevolle rol kan speel om gebruikers vertrouwd te maak met die onderskeie gebruike van die verskillende woordpare. Ook hier word 'n vak-interne geïntegreerde mikrostruktuur weer eens benut.

4.2.1.4 Ander data in die semantiese subkommentaar

Ruimtebesparing is geen groot kwessie in aanlyn woordeboeke nie. Daar is voldoende bergings- en aanbiedingsruimte om afstand te doen van ruimtebesparingsmetodes soos teksverdichting asook makrostrukturele nesting en nisting. Die vormkommentaar van 'n woordeboekartikel kan 'n gleuf bevat waarin verskillende tipes ongelede woorde verstrek word ter illustrasie van woordboumoontlikhede met die woord wat deur die lemma verteenwoordig word as boumiddel. Sulke gelede woorde word om morfologiese redes opgeneem en nie om as onverklaarde lemmas implisiete betekenisleiding te gee nie. As deel van die semantiese opdrag van 'n aanlyn woordeboek kan nisting en nesting uitgeskakel word ten gunste van 'n opname van daardie betrokke gelede leksikale items wat as lemmas kwalifiseer, as gidselemente van artikels wat 'n volledige bewerking bied. Hier is dan 'n eksplisiete oordrag van die gelede woorde se betekeniswaarde en nie, soos in die geval van onverklaarde lemmas, slegs 'n implisiete aanduiding nie. In aanlyn woordeboeke behoort eksplisiete betekenisoordrag prioriteitswaarde te hê.

Die aanbod van semantiese data in verklarende woordeboeke kan nog veel verder uitgebrei word en die vernuwende benutting van leksikografiese strukture dra daartoe by om 'n geskikte huisvesting van data te bewerkstellig. Woordeboeke is nie meer net draers van tekstipes nie, vergelyk Wiegand (1996), maar hulle bevat ook nietekstuele data soos klank- en beeldmateriaal. Sulke data kan ook as deel van die semantiese kommentaar aangebied word. In die horisontale ordening van data kan 'n semantiese subkommentaar 'n vak bevat met illustrasieprente wat in 'n verhouding van ostensiewe adressering (Gouws 1994) tot die lemma staan.

Aanlyn woordeboeke het verskillende soekposisies, te wete:

- die *soeksone*, dit is 'n gleuf in 'n woordeboekartikel waar 'n bepaalde aanduiding of datatipe geplaas word;

- die *soekgebied*, dit is die individuele woordeboekartikel;
- die *soekveld*, dit is die woordeboek se sentrale teks wat die verskillende artikeltrajekte bevat;
- die *soekstreek*, dit is die volle woordeboek met sy sentrale teks plus die verskillende buitetekste;
- die *soekdomein*, dit is die portaal waarbinne aanlyn woordeboeke geplaas kan word; en
- die *soekuniversum*, dit is die omgewing buite die soekdomein, byvoorbeeld die internet of ander woordeboekportaal-eksterne bronne waartoe gebruikers via die woordeboek toegang het (Gouws 2023: 383, e.v.)

'n Optimale benutting van soekposisies kan ook die uitbreiding van die aanbod van semantiese data bevoordeel. Hier is dit veral die soekdomein en die soekuniversum wat nuwe moontlikhede skep.

4.3 Semantiese data vanuit die soekdomein en die soekuniversum

Die soeksone, soekgebied, soekveld en soekstreek is tradisionele soekposisies binne 'n woordeboek en terreine waar semantiese data geplaas kan word — soos gesien in die besetting van vlakke en vakke in woordeboekartikels, dit is in 'n woordeboek se soekgebiede. Die soekdomein en soekuniversum is soekposisies buite die woordeboek wat as vertrekpunt van 'n woordeboekraadplegingsprosedure dien, maar wat 'n verrykingswaarde aan 'n woordeboek bied wat die aard en omvang van onttrekbare inligting betref.

Daar is reeds vroeër in hierdie artikel verwys na Bergenholtz en Bergenholtz (2017) se voorstelle om meerdere woordeboeke aan een databasis te onttrek. Met behulp van slim sagteware kan leksikografiese data ook aan sowel woordeboeke as ander bronne onttrek word. Die leksikograwe van 'n verklarende aanlyn woordeboek wat in 'n woordeboekportaal optree, moet bewus wees van die tipe inhoud van die ander woordeboeke in die portaal. 'n Woordeboekportaal-interne kruisverwysingstelsel kan dit dan moontlik maak om vanuit 'n artikel in een woordeboek na 'n verlangde datatipe in 'n ander woordeboek in dieselfde soekdomein te beweeg. 'n Tipe semantiese data wat nie in 'n gegewe woordeboek aangebied word nie kan vanuit daardie woordeboek in 'n ander woordeboek van die portaal gevind word. Wisselwerking binne 'n soekdomein kan ook tot 'n omvattender beskikbaarstelling van semantiese data lei.

Die soekuniversum maak dit moontlik om toegang te kry tot data wat buite die grense van 'n woordeboek en woordeboekportaal val. Deur die gebruik van 'n leksikografiese datatrekkingstruktuur, vergelyk Gouws (2018), is die onttrekking van inligting aan data uit leksikografiese en ander bronne moontlik. Die inligting wat onttrek word aan die data in 'n koteksvak van 'n woordeboekartikel kan aangevul word met data vanuit die woordeboek se korpus. Hier moet die gebruiker daarop bedag wees dat die korpusdata nie geredigeer is nie. Bykomende korpusdata kan bydra tot hoër vlakke van teksbegrip en tot 'n beter interpretasie van die gebruik van 'n bepaalde woord. Dit bied ook meer insig in die

breër konteks van 'n sitaat. So 'n datatrekking brei die omvang van die betekenis- en gebruiksleiding uit wat 'n woordeboek bied en verhoog die status van die woordeboek as 'n gesaghebbende en verteenwoordigende bron van taalkundige kennis.

'n Datatrekkingstruktuur skep ook die moontlikheid om vanuit 'n woordeboekartikel inligting aan internetbronne te onttrek. Veral ensiklopediese inligting ter bevrediging van 'n woordeboek se kognitiewe funksie kan op hierdie manier verkry word. Hierdie soort datatrekking is reeds 'n keuse in verskeie woordeboeke. Vergelyk in hierdie verband figuur 14, 'n skermkoot van 'n deelartikel van die lemma *auto* in *dict.cc*. 'n Klik op die inligtingsmerker naas 'n bepaalde woord bied 'n opskietspyskaart van waar gebruikers deur 'n verdere klik na onder meer Google gelei kan word om toegang te kry tot aanvullende semantiese en ander data:

English	German
NOUN an auto autos SYNO auto automobile car machine ...	NOUN das Auto die Autos SYNO auto ... selbst... [alles] Auto ...
auto-aggressive {adj} psych.	autoaggressiv
auto-centered {adj} {past-p} [Am.]	autozentriert
auto-centred {adj} {past-p} [Br.]	autozentriert
auto-discriminative {adj} comp.	selbsterkennend

Nouns	Frequency
car	5824
automobile [esp. Am.]	95
auto [Am.] [coll.] [automobile]	51
motor (US, UK, Aust.)	35
Personenwagen (m)	29
Auto (n)	17
Personenkraftwagen (m) <PKW, Pkw>	13
Anrufbeantwortungsmodus (m)	
Auspuffgase (pl)	
Auspuffgase (pl)	
Autoaggression (f)	
Autoaufbereitung (f)	
Autobiografie (f)	
Autobranche (f)	
Autodafé (n)	
Autodiebstahl (m)	
Autoerotik (f)	
Autofähre (f)	
Autohändler (m)	
Autohändlerin (f)	
Autohersteller (m)	
Autohersteller (pl)	
Autoimmunerkrankung (f)	
Autoimmunkrankheit (f)	

Figuur 14: Uit *dict.cc*

5. Ten slotte

Die aanbieding van semantiese data is een van die hoofopdragte van die opstellers van verklarende woordeboeke. Die moontlikhede wat die aanlyn omgewing bied, kan daartoe lei dat woordeboeke 'n veel omvattender en uitvoeriger aanbod van semantiese data vertoon. Leksikograwe moet op 'n kundige en dikwels vernuwende manier met woordeboekstrukture omgaan. Daar moet weg-beweeg word van 'n dataverspreiding en -plasing wat alle data gelyktydig sigbaar maak vir gebruikers. Deur data in verskillende vlakke en vakke aan te bied en volgens 'n veelfasige geïntegreerde mikrostruktuur te orden, kan gebruikers slegs daardie semantiese inligting onttrek waaraan hulle 'n behoefte het — sonder om deur al die soeksones in 'n omvangryke woordeboekartikel te worstel. Woordeboek-interne data kan aangevul word deur aan gebruikers toegang tot data in die soekuniversum te gee. Sulke prosedures verhoog die waarde van 'n woordeboek as 'n omvattende bron van taalkundige kennis.

Om woordeboeke relevant te hou vir die teikengebruikers moet leksikograwe pogings aanwend om meer en meer gerigte data in hulle woordeboeke op te neem, maar ook om die woordeboek as vertrekpunt vir soektogte in ander bronne te laat dien. Hiervoor skep die dekking van semantiese data 'n ideale geleentheid.

6. Bronnelys

6.1 Woordeboeke

ANW = *Algemeen Nederlands Woordenboek*. <http://anw.inl.nl/> [April 2025]

De Villiers, M. et al. (Reds.). 1988⁷. *Nasionale woordeboek*. Kaapstad: Nasou.

dict.cc = <https://www.dict.cc/> [Mei 2025].

elexiko = <http://www.owid.de/wb/elexiko/start.html> [April 2025]

HAT = Luther, J. et al. (Reds.). 2015⁶. *HAT: Handwoordeboek van die Afrikaanse Taal*. Kaapstad: Pearson.

Louw, J.P. en E. Nida. 1988. *Greek-English Lexicon of the New Testament Based on Semantic Domains*. New York: United Bible Societies.

Martin, W. et al. (Reds.). 2011. *Prisma Groot woordenboek Afrikaans en Nederlands*. Houten: Prisma.

WAT = *Woordeboek van die Afrikaanse Taal*.

6.2 Ander bronne

Bergenholtz, I. en H. Bergenholtz. 2017. One Database, Four Monofunctional Dictionaries. *Hermes* 50: 119-125.

Bergenholtz, H., T.J.D. Bothma en R.H. Gouws. 2011. A Model for Integrated Dictionaries of Fixed Expressions. Kosen, I. en K. Kosen (Reds.). 2011. *Electronic Lexicography in the 21st Century: New Applications for New Users: Proceedings of eLex 2011, Bled, Slovenia, 10-12 November 2011*: 34-42. Ljubljana: Trojina, Institute for Applied Slovene Studies.

Combrink, J.G.H. 1979. Die sesde deel van die WAT. *Standpunte* 140 (32-2): 49-64.

- DiMuccio-Failla, P.V.** 2025. A Theory for a Usage-Based Cognitive Lexicography. Giacomini, L. en V. Piuino (Reds.). 2025. *Patterns of Meaning in Lexicography and Lexicology*: 19-90. Berlyn/Boston: De Gruyter.
- Fodor, J.D.** 1977. *Semantics: Theories of Meaning in Generative Grammar*. New York: Crowell.
- Gouws, R.H.** 1987. Lexical Meaning versus Contextual Evidence in Dictionary Articles. *Dictionaries* 9: 87-96.
- Gouws, R.H.** 1989. *Leksikografie*. Kaapstad: Academica.
- Gouws, R.H.** 1994. Ostensiewe adressering in vertalende woordeboeke. *Lexikos* 4: 61-85.
- Gouws, R.H.** 2003. Aspekte van mikrostrukturele verskeidenheid en inkonsekwentheid in woordeboeke. *Lexikos* 13: 92-110.
- Gouws, R.H.** 2006. Die zweisprachige Lexikographie Afrikaans–Deutsch — Eine metalexikographische Herausforderung. Dimova, A., V. Jesenšek en P. Petkov (Reds.). 2006. *Zweisprachige Lexikographie und Deutsch als Fremdsprache*: 49-58. Hildesheim: Georg Olms.
- Gouws, R.H.** 2018. 'n Leksikografiese datatrekkingstruktuur vir aanlyn woordeboeke. *Lexikos* 28: 177-195.
- Gouws, R.H.** 2023. Nuwe raamstruktuurtypes in aanlyn woordeboeke. *Lexikos* 33(1): 382-403.
- Gouws, R.H.** (om te verskyn). Definitional Methods in Lexical Semantics.
- Gouws, R.H. en T.J.D. Bothma.** 2024. Aspekte van inligtingsonttrekkingstrukture in aanlyn woordeboeke. *Lexikos* 34(1): 355-370.
- Gove, P.B.** 1966. Self-Explanatory Words. *American Speech* 41: 182-198.
- Hausmann, F.J. et al. (Reds.).** 1989–1991. *Wörterbücher. Dictionaries. Dictionnaires. An International Encyclopedia of Lexicography*. Berlyn: De Gruyter.
- Klotz, M. en T. Herbst.** 2016. *English Dictionaries. A Linguistic Introduction*. Berlyn: Erich Schmidt.
- Levisen, C. en S. Waters (Reds.).** 2017. *Cultural Keywords in Discourse*. Amsterdam: John Benjamins.
- Louw, J.P.** 1985. A Semantic Domain Approach to Lexicography. Louw, J.P. (Red.). 1985. *Lexicography and Translation*: 157-197. Kaapstad: Bybelgenootskap van Suid-Afrika.
- Lyons, J.** 1977. *Semantics*. Cambridge: Cambridge University Press.
- Müller-Spitzer, C. en A. Kopenig.** 2014. Questions of Design. Müller-Spitzer, C. (Red.). 2016. *Using Online Dictionaries*: 189-204. Berlyn/Boston: De Gruyter.
- Odendal, F.F.** 1979. Plus positief en plus negatief. *Tydskrif vir Geesteswetenskappe* 19(1): 24-41.
- Scerba, L.** 1940. Opyt obshchei teorii leksikografii. *Izvestiia Akademii Nauk SSSR, Otdelenie literatury i iazyka* 3: 89-117. (Engelse vertaling deur D.M.T.Cr. Farina 1995: Towards a General Theory of Lexicography. *International Journal of Lexicography* 8(4): 314-350).
- Tarp, S.** 2009. Beyond Lexicography: New Visions and Challenges in the Information Age. Bergenholtz, H., S. Nielsen en S. Tarp. (Reds.). 2009. *Lexicography at a Crossroads: Dictionaries and Encyclopedias Today, Lexicographical Tools Tomorrow*: 17-32. Bern: Peter Lang.
- Wiegand, H.E.** 1989a. Arten von Mikrostrukturen im allgemeinen einsprachigen Wörterbuch. Hausmann, F.J. et al. (Reds.). 1989–1991: 462-501.
- Wiegand, H.E.** 1989b. Aspekte der Makrostruktur im allgemeinen einsprachigen Wörterbuch: alphabetische Anordnungsformen und ihre Probleme. Hausmann, F.J. et al. (Reds.) 1989–1991: 371-409.
- Wiegand, H.E.** 1996. A Theory of Lexicographic Texts. An Overview. *SA Journal of Linguistics* 14(4): 134-149.
- Wiegand, H.E. en M. Smit.** 2013. Microstructures in Printed Dictionaries. Gouws, R.H. et al. (Reds.). 2013. *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*: 149-214. Berlyn/New York: De Gruyter.

<http://lexikos.journals.ac.za>; <https://doi.org/10.5788/35-2-2108> (Volle uitgawe / Full volume)
<https://lexikos.journals.ac.za>; <https://doi.org/10.5788/35-2-2075> (Article)

74 Rufus H. Gouws

Wierzbicka, A. 1972. *Semantic Primitives*. Frankfurt: Athenäum.

Wittgenstein, L. 1953. *Philosophical Investigations*. Oxford: Blackwell.

Zgusta, L. 1971. *Manual of Lexicography*. Den Haag: Mouton.

Lexicographic Description of a Polysemous Word in a Learner's Dictionary Based on Its Lexical Prototype

Anton Karasev, *Département d'Anglais, Laboratoire Ligérien
de Linguistique (VETD-DO), Université de Tours, Tours, France*
(anton.karasev@univ-tours.fr)
(<https://orcid.org/0009-0005-7311-6922>)

Abstract: This paper presents a practical application of the lexical prototype theory to lexicographic practice, proposing a draft model of a learner's dictionary entry for a polysemous word based on its lexical prototype. The lexical prototype (LP) is regarded as the semantic core of a polysemous word that functions at the level of the language system and is represented in actual speech by contextual meanings: the primary meaning (PM), metonymic, and metaphorical senses. Drafting a dictionary entry based on the LP requires reconstructing the LP through analysing the contextual meanings in relation to the PM in order to establish the minimal set of semantic features that are shared by all the senses of the word. The proposed method of lexicographic description of polysemous words offers a new perspective on polysemy as a linguistic phenomenon as well as moves away from the list-based method.

Keywords: LEXICOGRAPHY, LEARNER'S DICTIONARY, DICTIONARY ENTRY, POLYSEMY, LEXICAL PROTOTYPE, PRIMARY MEANING, METONYMY, METAPHOR, COGNITIVE LINGUISTICS

Opsomming: Die leksikografiese beskrywing van 'n polisemiese woord in 'n aanleerderswoordeboek gebaseer op die leksikale prototipe daarvan. In hierdie artikel word 'n praktiese toepassing van die leksikale prototipe-teorie in die leksikografiese praktyk aangebied en 'n konsepmodel van 'n aanleerderswoordeboekinskrywing vir 'n polisemiese woord gebaseer op die leksikale prototipe daarvan, word voorgestel. Die leksikale prototipe (LP) word beskou as die semantiese kern van 'n polisemiese woord wat op die vlak van die taalsisteem funksioneer en word in werklike spraak deur kontekstuele betekenisse verteenwoordig: die primêre betekenis (PB), metonimiese, en metaforiese betekenisse. Die opstel van 'n woordeboekinskrywing wat op die LP gebaseer is, verg die rekonstruksie van die LP deur die analisering van die kontekstuele betekenisse in verhouding tot die PM om sodoende die minimale stel semantiese kenmerke wat deur al die betekenisse van die woord gedeel word, te bepaal. Die voorgestelde metode vir die leksikografiese beskrywing van polisemiese woorde bied 'n nuwe perspektief op polisemie as 'n linguïstiese verskynsel en beweeg terselfdertyd weg van die lysgebaseerde metode.

Sleutelwoorde: LEKSIKOGRAFIE, AANLEERDERSWOORDEBOEK, WOORDEBOEKINSKRY-

WING, POLISEMIE, LEKSIKALE PROTOTIPE, PRIMÈRE BETEKENIS, METONIMIE, METAFOR, KOGNITIEWE LINGUISTIEK

Introduction

When making a dictionary, a lexicographer faces a difficult task: on the one hand, the work must be theoretically grounded, on the other hand, it is supposed to satisfy the needs of the target audience, which determines the popularity of the dictionary and keeps it in demand (Slonimskaya 1990: 49). As a rule, dictionary users are reluctant to choose what they are not accustomed to. As a result, the structure and content of a lexicographic work are shaped as much by the expectations of the target audience as by linguistic factors. Therefore, dictionary-making is most often built on the achievements of previous generations of lexicographers both in compiling the lexicon and in writing dictionary entries (Jackson 2002: 27). Thus, the current consensus is that polysemy must be presented in the dictionary in a linguistically justified manner (Lodwig and Barret 1967: 53-54). Despite significant progress in the field of cognitive semantics, the methods of lexicographic description of lexical meanings have not changed much over time (Karasev 2014a; Halas 2016: 125). Nearly a century after the first learner's dictionaries came into being — which once stood at the forefront of both practical and theoretical lexicography (Cowie 2007: 14; Jackson 2002: 129) — no significant progress has been made in this field. For example, presenting polysemy as a list of numbered definitions introduced by Palmer has remained largely unchanged since the 1930s (Cowie 2007: 44). This is how practising lexicographers contribute to the development of linguistics: the methods they use to represent the semantic structure of words give rise to discussions whether dictionary definitions correspond to the actual lexical meaning (Arnold 2010: 108). There have been attempts to create lexicographic descriptions based on prototype theory (Halas 2016), and to describe meanings in a way that reflects their mental representations (Rivelis 2009). Opponents of such attempts argue that since we cannot know exactly how meanings are stored in the mind, it is impossible to describe them in the same way a dictionary. If this were to be discovered, the verbal description of meanings would differ significantly from how those meanings are represented in the mind of a native speaker (Zaliznyak 2004). Nevertheless, such an attempt still seems justified and, especially in the field of learner's lexicography, as it may be useful to reconstruct the systemic meaning before examining actual meanings (Guillaume 1973: 143).

This paper proposes a theoretically sound model for a learner's dictionary entry of a polysemous word based on lexical prototype theory (Arkhipov 2003: 79; Levchina 2003: 269; Pesina 1998, 2005a, 2005b; Shiryayeva 2008), developed in the works of Arkhipov and his followers. According to this theory, a word's meaning within the language system reflects a generalisation of its various uses in everyday speech. Grasping this generalised meaning can be especially helpful for

foreign language learners, as it allows them to recognise connections between specific contextual meanings, or senses, that might otherwise be unclear.

Polysemy in learner's dictionaries

The value of a dictionary lies in its ability to present the linguistic information in a visible, complete diverse manner. It serves as an essential tool of understanding language phenomena as it influences both their perception and, to some extent, linguistic consciousness. One might as well speak of a special "dictionary view" of reality (Devkin 2000: 10, 14-15; Atkins and Rundell 2008: 2). At the same time, a dictionary — especially a learner's dictionary — serves as an authority for the majority of its users. Therefore, it is essential to develop lexicographic principles that ensure descriptions of linguistic phenomena are as accurate and faithful to linguistic reality as possible. In this regard, an important issue is the correspondence between a word's meaning and its dictionary definition. Specifically, it is necessary to understand whether the definition reflects the actual lexical meaning, the reference underlying that meaning, or whether it merely provides a verbal description of a real-world object (Ufimtseva 1984: 135). Another key question concerns the extent to which the lexicographic treatment of polysemous words in dictionaries aligns with linguistic consciousness.

In theoretical literature, as well as in lexicography textbooks, the issue of polysemy and its lexicographic representation is usually described as resolved (Apresyan 1984: 409; Jackson 2002: 88; Dubchinskiy 1994: 10, 24). In these works, the authors outline the existing ways of presenting word meanings in dictionaries. Other researchers describe it as problematic (Hanks 2000: 211–212), though they don't propose specific solutions. Interestingly, the idea of polysemy as existence of several independent, though related meanings of a word has been shaped by dictionary making practice (Apresyan 1984: 186; Katsnelson 2011: 40; Mednikova 2010: 17). This leads to a vicious circle: a practical lexicographer finds theoretical justification for the list-based method of describing polysemy in the linguistic literature, while a theoretical linguist uses dictionary data as a basis for describing the phenomenon of lexical polysemy. It is also interesting to note that English-language lexicography does not seek to establish or highlight the closeness of separate senses of a word within dictionary entries (Litvin 1984: 5-6); instead, more attention is paid to reflecting the relation between lexical units (Soldatov 1997: 120). This leads to the fact that the meanings of polysemous words in monolingual learner's dictionaries of English are often presented as lengthy lists of definitions.

Corpus studies confirm that the meanings of a polysemous word do not exist as 'checklists': "The numbered lists of definitions found in dictionaries have helped to create a false picture of what really happens when language is used." (Hanks 2000: 205). Since human memory doesn't have the capacity to store every variation of the meanings of all known words (Brudny 1971: 19), the information about language units is retained in an individual's memory not as extended

dictionary definitions, but in a more compact form, together with the mechanisms that enable the actualisation of meanings in context (Pertsov 2001: 43; 22). Thus, a dictionary presents a purely theoretical scheme which reflects the lexicographer's interpretation of the totality of actualisations of language-system units in speech, created through the "cataloguing" of meanings attested in texts (Arkhipov 2004: 79). This scheme does not represent the actual meaning of the word itself, but rather its "meaning potentials" that is, "potential contributions to the meanings of texts and conversations in which the words are used, and activated by the speaker who uses them" (Hanks 2000: 211). In this sense, the dictionary constitutes a minimised model of the speech usage of language units and does not provide a reliable representation of a word's semantics (Chudinov 2003: 34).

This means that the word meaning at the level of the language system must differ significantly from its realisation in speech. Given that human memory is not "a warehouse of individual traces duplicating the events of the external world, but a creative process of constructing these events" (Zalevskaya 2000: 53), the meaning at the language-system level, regardless of the number of specific meanings recorded in the dictionary, must be sufficiently compact and abstract. Each contextual meaning correlates with the systemic meaning in the same way that an allophone or allomorph relates to a phoneme or morpheme, respectively (Pesina 2005c: 76; Jakobson 1985: 177). The idea of "semantic compactness," as well as the distinction between invariants at the level of language and variants at the level of speech — originally distinguished in phonology and later in morphology — has led to the emergence of hypotheses about the existence of a core meaning, or an invariant meaning of a word at the level of the language system (Arkhipov 2003; Budagov 1947; Shmelev 1973; Smirnitsky 1954: 23; Vinogradov 1947: 34).

This systemic meaning of a word is contrasted with its contextual meanings, or senses. The systemic meaning can be established out of context, while contextual, or derived, meanings are determined only through actual use (Kuryłowicz 1955: 73-81). The systemic meaning contains a certain feature that links various senses of a word and serves as the basis for new derived meanings (Zvegintsev 1957: 205-207). Despite the existence of such studies, their results, with rare exceptions (Rivelis 2009), are not reflected in lexicography.

Polysemy and lexical prototype

Cognitive linguistics, and in particular prototype theory, offers a perspective on polysemy that is more consistent with linguistic consciousness: it views polysemy as the presence of several interrelated meanings within a word, characterized by gradedness, fuzziness, flexibility, and clustering of senses (Geeraerts 2006: 141; Nikitin 2008: 71), all organized around a prototype — the best representative of the semantic structure (Rosch 1978: 54; Lakoff 1984). Among those who support the existence of a general, systemic meaning for polysemous words, prototype

theory has led to the hypothesis that such a best representative may be so effective that it can serve as the invariant meaning (Pesina 2005c: 65). I. Arkhipov proposed calling this semantic core the lexical prototype (LP) (Arkhipov 1998: 15). The LP represents a minimal set of semantic features shared by all the senses of the word (Leshcheva 1985: 5). It includes both the essential features associated with the form of the linguistic sign and the results of semantic derivation (Pesina 2005a: 19). Therefore, the LP serves, first, as a sufficient condition for interpreting a word's actual meaning (Pesina 1998: 6; Levchina 2003: 8), and second, as the foundation for the formation of all the meanings of a polysemous word (Pesina 2005c: 79).

Research on polysemy is closely associated with the study of lexical-semantic variation, according to which a same lexeme is represented in different contexts as its lexico-semantic variants (LSVs) (Ufimtseva 1962: 91; Ufimtseva 2010: 31; Yakovlyuk 2009: 156-157). In this framework, the LP operates as the invariant unit at the language-system level (systemic meaning), while each LSV represents its realisation at the level of speech (actual meaning) reflecting the natural mechanisms of semiosis (Karasev 2014c).

The concept of LP has practical applications in lexicography. For example, Arkhipov, Novikov and Pesina (2006) suggest placing the LP at the beginning of a dictionary entry as a generalised systemic meaning of a polysemous word that serves as a motivational basis for all the senses listed below. This dictionary entry is, however, list-based and the LP serves only a supplementary function. Independently of research on the LP, a similar model of a dictionary entry was proposed in (Halas 2016), where a lexicographic description of a polysemous word is based on the radial set model with reference to the prototypical meaning. This entry constitutes a complex hierarchical structure that demonstrates the semantic relationships between the word's senses. Nevertheless, it still does not depart from the traditional list-based method.

As noted above, the LP represents the minimal set of the most essential features sufficient for the adequate interpretation of the contextual meaning (Arnold 2010: 8), and is likely to be the systemic meaning of the word that enables its functioning in speech (Arutyunova 1999: 269).

Within the framework of LP theory, a polysemous word has three types of meanings: the primary meaning (PM), metonymic meanings, and metaphorical meanings. The PM also corresponds to the prototypical meaning in the theory of Brugman and Lakoff (Brugman and Lakoff 1988). The semes underlying metonymic and metaphorical transfers are part of the LP. Since, in real communicative situations, actual meanings are derived based on context, a dictionary entry structured around the LP includes the LP itself as well as examples of the headword in various contexts, accompanied by prompts — similar to signposts in LDOCE — that enable the dictionary user on their own to derive the meaning of each LSV with reference to the LP.

The systemic meaning exists as a part of the cognitive structure of a language speaker, and therefore is hidden from direct access. However, it can be estab-

lished based on the actual meanings of a word that are catalogued in existing dictionaries (Alefirenko 2005: 110). The LP can be described as the minimal semantic core of the word meaning, which functions at the level of the language system and which can serve as a basis for the construction of a consistent semantic structure of a word (Pesina 1998: 8).

Dictionary definitions provide detailed descriptions of contextual meanings, or LSVs, which offer valuable insight into how language speakers understand meanings, they are especially useful for establishing the LP.

The procedure for establishing the LP involves a componential analysis of dictionary definitions and an interpretation of contextual meanings to identify the semantic components that constitute the features common to all senses. This type of analysis, applied to nouns, verbs, and adjectives, is set out in detail in the dissertation (Karasev 2014a) as well as in related papers (Karasev 2013, 2014b), with no apparent variation across parts of speech. This paper proposes a draft model of a learner's dictionary entry for the verb *see* based on its LP.

Procedure for establishing the LP

To compile a dictionary entry for a polysemous word based on its LP, the first step is to formulate the LP of this verb. There are two stages of the procedure:

1. Identification of the PM of the word through the analysis of definitions from several dictionaries;
2. Interpretation of the word's LSVs based on the PM, in order to examine how the word's senses relate to PM and determine the minimal set of semantic features that constitute the LP.

Below are the definitions of the primary meaning of the verb *see* from seven general-purpose and two learner's dictionaries.

General-purpose dictionaries

<i>Cambridge Dictionary</i>	to be conscious of what is around you by using your eyes
<i>Oxford English Dictionary</i>	to perceive with the eyes
<i>The Chambers Dictionary</i>	to perceive by the sense operated in the eyes
<i>American Heritage Dictionary</i>	to perceive with the eye
<i>Merriam-Webster Dictionary</i>	to perceive by the eye
<i>Webster's Online Dictionary</i>	perceive by sight or have the power to perceive by sight

Learner's dictionaries

<i>Collins COBUILD</i>	when you see something, you notice it using your eyes.
<i>Longman Dictionary of Contemporary English</i>	to notice or examine someone or something, using your eyes
<i>Oxford Advanced Learner's Dictionary</i>	to become aware of somebody / something by using your eyes

As shown in the tables above, these definitions consistently feature, in various forms, the same semantic component [EYE PERCEPTION]: *perceive with the eye(s)/by sight/by the sense operated in the eyes*. Thus, the PM of the verb *see* can be formulated as: *perceive with the eyes*. The determinant may be adapted to the context if necessary. All the sentences containing the LSVs of the verb *see* below were taken from the British National Corpus. The selection of contexts roughly reflects the meanings of the verb *see* traditionally included in dictionaries.

To further identify the LP, the verb's LSVs are analysed against its PM. For this purpose, the contexts in which the LSVs of *see* are used were examined in detail.

LSV in context	PM	Interpretation
(1) If the operation is successful, he will be able to <i>see</i> again	<i>perceive with the eyes</i>	If the operation is successful, he will be able to <i>perceive with the eyes</i> again
(2) Can you <i>see the</i> bird in that tree?	<i>perceive with the eyes</i>	Can you <i>perceive</i> the bird in that tree <i>with your eyes</i> ?

In contexts (1) and (2), the verb *see* is used in its PM. In (1), the verb is intransitive and means *have the power to perceive by sight* (MWD), whereas in (2) it is transitive and means *have the power to perceive by sight*. Since seeing objects is possible only if one has *the power of sight* and to *have the power to perceive by sight* implies the ability to *see* objects, this distinction is irrelevant for the purposes of this study. Therefore, the same meaning is realised in both cases and it corresponds to the PM: *see*₁ — *perceive with the eyes*.

(3) Did you <i>see</i> that documentary on Channel 4 last night?	<i>perceive with the eyes</i>	Did you <i>perceive</i> that documentary on Channel 4 <i>with your eyes</i> last night?
--	-------------------------------	---

Context (3) suggests a purposeful action directed at *the documentary* as its object: *obtain the content of the documentary by perceiving it with the eye*. This action is based on the physical ability to see, but it is not limited to it and involves the

interpretation of what has been seen. The sense of the verb *see* in the analysed context is metonymic, because the interplay of concepts occurs within the same domain: the EYE PERCEPTION. This is the case of a metonymic transfer: sensory perception — obtaining sensory information as the goal of perception (Ageeva 1990: 10). This contextual meaning of *see* can be phrased as *see₂* — *watch something (film, television, etc.) by perceiving with it the eyes*.

The same mechanism works in contexts (4) and (5):

- | | | |
|---------------------------------------|-------------------------------|---|
| (4) Please <i>see</i> who's knocking. | <i>perceive with the eyes</i> | Please <i>perceive</i> who's knocking <i>with your eyes</i> . |
| (5) <i>See</i> chapter 12. | <i>perceive with the eyes</i> | <i>Perceive</i> chapter 12 <i>with your eyes</i> . |

Unlike *see₂*, the focus in (4) and (5) is on the processing of information: looking at something → perceiving it with the eyes → receiving the information. The same meaning is found in context (6).

- | | | |
|---|-------------------------------|--|
| (6) I must <i>see</i> your passport before you can enter the country. | <i>perceive with the eyes</i> | I must <i>perceive</i> your passport <i>with my eyes</i> before you can enter the country. |
|---|-------------------------------|--|

This context doesn't imply a requirement to physically observe the passport, but rather to examine its content, that is to look at it in order to obtain information. Since the verb *see* refers to a purposeful action based on visual perception, this is a metonymic sense that can be presented as follows: *see₃* — *look up, refer to, or find something out by perceiving it with the eyes*.

- | | | |
|---|-------------------------------|--|
| (7) We still <i>see</i> each other a couple of times a month. | <i>perceive with the eyes</i> | We still <i>perceive with</i> each other <i>with our eyes</i> a couple of times a month. |
|---|-------------------------------|--|

In (7), the verb *see* is also used in a metonymic sense, as it does not describe eye perception itself, but rather a certain kind of interaction that requires visual perception. This transfer is based on the idea of meeting someone one knows socially or by chance while *perceiving them with the eyes*, therefore: *see₄* — *visit or meet someone one knows socially or by chance while perceiving them with the eyes*.

In a similar sense, the verb *see* is used in contexts (8) and (9).

- | | | |
|--|-------------------------------|--|
| (8) Can I <i>see</i> you home? | <i>perceive with the eyes</i> | Can I <i>accompany</i> you and <i>perceive with my eyes</i> that you are at home? |
| (9) My mother used to <i>see</i> me across the road. | <i>perceive with the eyes</i> | My mother used to <i>go with</i> me across the road <i>while perceiving me with her eyes</i> |

What makes the verb *see* special is that visual perception is the primary means of obtaining information about the world (Arutyunova 1999: 11; Dretske 1969: 77). That is why it can be used to denote a range of actions that somehow involve visual perception. In contexts (8) and (9), the verb *see* refers to the idea of accompanying someone somewhere and making sure that they are there *by perceiving it with your eyes*. In real communication, speakers tend to choose the most economical way of expressing this complex idea using a single word *see*. The combinability of the verb — *see someone somewhere* — hints at this concept. Thus, the verb *see* is used metonymically in a sense directly related to visual perception: *see₅ — accompany somebody to make sure they get there by perceiving it with the eyes*.

- | | | |
|--|-------------------------------|--|
| (10) Please <i>see</i> that the lights are switched off before you leave. | <i>perceive with the eyes</i> | Please <i>make sure that</i> the lights are switched off before you leave [AS IF] <i>by perceiving it with your eyes</i> . |
| (11) Could you <i>see</i> that the kids are in bed by 8 o'clock? | <i>perceive with the eyes</i> | Could you make sure that the kids are in bed by 8 o'clock [AS IF] <i>by perceiving it with your eyes?</i> |
| (12) I have had them checked to <i>see</i> that they are sound, and they are no danger to any structure. | <i>perceive with the eyes</i> | I have had them checked to make sure that they are sound, and they are no danger to any structure [AS IF] <i>by perceiving it with my eyes</i> . |
| (13) It's up to you to <i>see</i> that the job's done properly. | <i>perceive with the eyes</i> | It's up to you to make sure that the job's done properly [AS IF] <i>by perceiving it with your eyes</i> . |

In contexts (10), (11), (12) and (13), the verb *see* is used in a metaphorical sense, as it is related to the PM through the comparative semantic feature [AS IF]. In forming this meaning, the source domain of EYE PERCEPTION is mapped onto the target domain MENTAL PERCEPTION, which yields the meaning *visually check that something is done*. It may seem that in contexts (10) and (11) the verb *see* is used literally, since the object of visual perception is, at first sight, physically visible objects: lights and kids. However, the position of the direct object of *see* is not taken by these nouns, but by object clauses, which are always associated with verbs of mental activity. This metaphorical sense of *see* can be formulated as follows: *see₆ — make sure that something is done as if by perceiving the result with the eyes*.

It is interesting to note in relation to the semantics of the verb *see* that, due to speakers' capacity of abstract thinking and imagination that stems from it, the boundary between verbally described, actually perceived, and mentally modulated objects is, to some extent, blurred. Thus, a native English speaker uses the verb *see* not only to refer to actually visible objects but also to imagined objects and situations. As the interpretation of meanings *see*₂, *see*₃, *see*₅ and *see*₆ demonstrates, visual perception is closely connected with the processing of information, which is why English speakers use this verb to refer to mental processes. The same is observed in context (14).

- (14) I can't *see* why he's so up- *perceive with the eyes* I can't *perceive*, [AS IF] *with*
set. *my eyes*, why he's so upset.

As in examples (10)–(13), in this case there is a contextual cue — an object clause as the direct object. Consequently, the contextual meaning of the verb *see* is associated with mental activity and belongs to the domain of MENTAL ACTIVITY or more precisely to its subdomain of INFORMATION PERCEPTION. Since the PM of the verb *see* belongs to the domain of EYE PERCEPTION, the sense found in context (14) is metaphorical. Visual perception is the basis of interaction with the world and the main source of information. Visual images become the foundation for logical reasoning, which is why the processes of visual perception and mental processing are closely related. In such situations, first of all, the [PERCEPTION] seme is actualised, promoting the metaphorical shift from EYE PERCEPTION to MENTAL PERCEPTION. As a result, the verb *see* is reinterpreted as a verb of intellectual activity. The resulting metaphorical sense can be worded as follows: *see*₇ — *understand something, as if by perceiving it with the eyes*.

- (15) I don't *see* the situation *perceive with the eyes* I don't *perceive* the situation
quite as negatively as you do. [AS IF] *with my eyes*, quite
as negatively as you do.

The sense of *see* in (15) is metaphorical, as it is derived from the PM on the basis of the comparison feature [AS IF]. This transfer results from the mapping of the EYE PERCEPTION domain onto the MENTAL PERCEPTION domain. The context provides the following cues. The object of perception is the abstract concept of *situation*, which is a mental construct. The adverb *negatively* expresses an attitude, which is not typical of sensory perception, but rather of cognitive information processing. As in the case of *see*₇ first of all the [PERCEPTION] feature is actualised. The mode of perception is determined by the interpreter based on these contextual cues. The difference between this meaning and *see*₇ is that in combination with the circumstance of the mode of action, the meaning of the verb becomes inseparably linked to the attitude expressed by the adverb *negatively*. The same is observed in example (16).

- (16) Having a child makes you *perceive with the eyes* *see* things differently. Having a child makes you *perceive* things differently [AS IF] *with the eyes*.

The adverb *differently* in (16) specifies the meaning of the verb *see*, thus serving as a contextual cue that facilitates the interpretation of its sense which can be presented as *see₈ — consider something in a particular way as if by perceiving it with the eyes*.

- (17) Can you really *see* her as the president? *perceive with the eyes* Can you really *perceive* her as the president [AS IF] *with the eyes*?

Contextual cues such as *really* and *as the president* indicate that the person referred to by the pronoun *she* does not hold a leading position in a certain institution. In context (17), as in (15) and (16), the same [PERCEPTION] is foregrounded, and this LSV of *see* is associated with intellectual activity. The same sense of *see* is found in examples (18) and (19).

- (18) He could *see* a great future for her in music. *perceive with the eyes* He could *perceive* a great future for her in music [AS IF] *with his eyes*.
- (19) She's got a new book coming out but I can't *see* it doing very well. *perceive with the eyes* She's got a new book coming out but I can't *perceive* it doing very well [AS IF] *with my eyes*.

Such contextual elements as *great future* and *can't see it doing very well* trigger an image of entities that do not exist in the physical world but are conceived of as real. Considering this, the sense of *see* in contexts (17), (18) and (19) can be described as *see₉ — imagine something which is not true as if by perceiving it with the eyes*.

- (20) As we *saw* in Chapter 2, the reasons for the war were complex [173]. *perceive with the eyes* As we *found out* from Chapter 2 [AS IF] we *perceived* this information *with our* eyes, the reasons for the war were complex.
- (21) I'll go and *see* what he wants [181]. *perceive with the eyes* I'll go and *find out* what he wants [AS IF] I'll *perceive* this information *with the eyes*.

In contexts (20) and (21), the verb *see* is used metaphorically: the domain of EYE PERCEPTION is mapped onto the domain of OBTAINING INFORMATION. The verb *see*

in this context does not refer to eye perception, but rather to the result of obtaining information. This sense is similar to *see*₃ and *see*₇. However, in (20) and (21), the verb is used to describe the result of thought processing of information obtained from a known source. The metaphorical transfer is based on likening visual perception to mental perception. This meaning can be presented as *see*₁₀ — *find something out as if through eye perception*.

- (22) He *saw* action in Vietnam. *perceive with the eyes* He *took part in* action in Vietnam [AS IF] *he perceived it with his eyes*.

The direct object of *see* in (22) refers to an abstract concept — *action*, which cannot be perceived visually. The metaphorical transfer is based on the mapping of the EYE PERCEPTION domain onto the EXPERIENCE domain: the idea of "seeing" is projected onto the whole perception of combat — *experience acquired as if through eye perception*. The same sense of *see* is found in (23):

- (23) The region has *seen* some of the fiercest fighting in the war. *perceive with the eyes* Some of the fiercest fighting in the war *took place* in the region [AS IF] *the region perceived it with the eyes*.

The semantic transfer in (23) is driven by the same mechanism: *experiencing something as if through eye perception*. In this case, a twofold understanding is possible. Either the noun *region* is used metonymically and refers to the people who live in that area, or it is used in its PM with the subject of perception being the region itself. Regardless of the interpretation, the meaning of the verb *see* remains metaphorical, as it is conditioned by the interaction of independent cognitive domains (cf. the interpretation in (22)). Thus, the meaning of the verb *see* in (22) and (23) takes the form of *see*₁₁ — *experience something as if by perceiving it with the eyes*.

The extensive polysemy of the verb *see* can likely be attributed to vision being the most dominant of the five senses and, consequently, essential for humans in navigating and interpreting their environment. Everything that falls within an observer's field of vision undergoes cognitive processing (Maturana 1980: 8). This is probably why all the LSVs of the verb *see* analysed above the include the integral feature [PERCEIVE].

If the verb refers to the information (cognitive change triggered by external causes) that is perceived visually the [WITH THE EYES] seme is present. The object of perception may be an abstract entity, as in contexts (16) and (23). In this case, the observer can only mentally "see" it and perceive the image if it were real (i.e., there is a change in its state caused by internal causes). For a human being whose nervous system is a closed neuronal network, there is no significant distinction between a visual image triggered by internal or external changes in relative neuronal activity (Ritchie 2013: 46). Therefore, actually seeing

with the eyes and "as if seeing with the eyes" is essentially the same process. That is why the seme [WITH THE EYES] is necessarily present in the LP formula.

Besides that, since the LP of a polysemous word underlies all of its senses, its formula must encompass both metonymic and metaphorical extensions. Since metonymic transfer is based on contiguity, a metaphorical sense of a verb may be described as action, process or state associated with the PN in some way. Unlike PM, all the metonymic senses of the verb *see* discussed above are not associated with a state, but rather with a purposeful action — obtaining information. Evidence for this is the use of these LSVs in the imperative mood (e.g., *Please see₃ who's knocking. See₃ chapter 12*) or in the Present Continuous (e.g., *Did you know that she is seeing₃ an older man.*)

As the analysis shows, all the LSVs of the verb *see* are motivated by its nominative and non-derivative meaning. The meaning of *see₁* is motivated on the basis of direct denotation, so in the dictionary entry it is formulated as:

LITERAL: *perceive with the eyes*

The meanings of LSVs *see₂–see₅* are metonymic extensions of the PM:

see₂ — *to watch (film, television, etc.) by perceiving with the eyes*

see₃ — *to look up, refer to, or find something out by perceiving it with the eyes*

see₄ — *to visit or meet someone one knows socially or by chance while perceiving them with the eyes*

see₅ — *to accompany somebody to make sure they get there by perceiving it with the eyes.*

Since the meanings of these LSVs are somehow associated with the PM through the idea of contiguity, the generalised formulation for the given senses is as follows:

METONYMIC: *to do something associated, in some way, with eye perception.*

The meanings of LSVs *see₆–see₁₁* are formed on the basis of metaphorical transfer:

see₆ — *make sure that something is done as if by perceiving the result with the eyes*

see₇ — *understand something as if by perceiving it with the eyes*

see₈ — *consider something in a particular way as if by perceiving it with the eyes*

see₉ — *imagine something which is not true as if by perceiving it with the eyes*

see₁₀ — *find something out as if through eye perception*

see₁₁ — *experience something as if by perceiving it with the eye*

The meanings of these LSVs include the comparison feature [AS IF], which associates it with the PM. In accordance with the conventions of LP theory, when used in the LP formula, this seme takes the form of *as it were*. The generalised metaphorical meaning of the above LSVs is formulated as follows:

METAPHORICAL: *perceive with the eyes, as it were*

Thus, the semantic structure of the verb *see* takes the form of:

LITERAL: *perceive with the eyes*

METONYMIC: *to do something associated, in some way, with eye perception*

METAPHORICAL: *perceive with the eyes, as it were.*

To formulate the LP all the three elements must be combined:

see — *perceive with the eyes, or do something associated, in some way, with eye perception, or as it were.*

This formula describes in the most general form all three types of meanings of the verb: the PM, metonymic and metaphorical senses. The following signposts are included in the dictionary entry to help infer contextual meanings:

*see*₂ — *watch (film, television, etc.) by perceiving with the eyes* → **WATCH (TELEVISION, ETC.)**

*see*₃ — *look up, refer to, or find something out by perceiving it with the eyes* → **CHECK/LOOK/REFER TO/FIND OUT**

*see*₄ — *visit or meet someone one knows socially or by chance while perceiving them with the eyes* → **MEET/VISIT**

*see*₅ — *accompany somebody to make sure they get there by perceiving it with the eyes* → **GO WITH SOMEBODY SOMEWHERE**

*see*₆ — *make sure that something is done as if by perceiving the result with the eyes* → **MAKE SURE**

*see*₇ — *understand something as if by perceiving it with the eyes* → **UNDERSTAND**

*see*₈ — *to consider something in a particular way as if by perceiving it with the eyes* → **CONSIDER**

*see*₉ — *imagine something which is not true as if by perceiving it with the eyes* → **IMAGINE**

*see*₁₀ — *find something out as if by eye perception* → **FIND OUT INFORMATION**

*see*₁₁ — *experience something as if by perceiving it with the eye* → **EXPERIENCE SOMETHING**

Model of a dictionary entry based on LP

Based on the analysis, a dictionary entry for the verb *see* is compiled:

LEXICAL PROTOTYPE: *perceive with the eyes, do something associated with eye perception, or as it were*

MEANINGS:

LITERAL: *perceive with the eyes.*

If the operation is successful, he will be able to see again. • Can you see the bird in that tree?

• The moment we saw the house, we knew we wanted to buy it.

METONYMIC: *do something associated, in a certain way, with eye perception*

<i>Did you see that documentary on Channel 4 last night?</i>	WATCH (TELEVISION, ETC.)
<i>Please see who's knocking</i> • Elements are usually classified as metals or non-metals (see chapter 12) (=refer to chapter 12 for further information). • <i>I must see your passport before you can enter the country.</i>	CHECK/LOOK/REFER TO/FIND OUT
<i>We still see each other a couple of times a month</i> • <i>See you at the station at 6 o'clock</i> • <i>See you</i> (informal, used to say goodbye).	MEET/VISIT
<i>Can I see you home?</i> • <i>My mother used to see me across the road.</i> • <i>Let me see you to the door.</i>	GO WITH SOMEONE SOMEWHERE
<i>Please see that the lights are switched off before you leave.</i>	MAKE SURE

METAPHORICAL: perceive with the eyes, as it were

<i>I can't see why he's so upset.</i> • <i>I see what you mean.</i> • <i>You see, the thing is, I'm really busy right now.</i>	UNDERSTAND
<i>Having a child makes you see things differently.</i> • <i>I don't see the situation quite as negatively as you do.</i>	CONSIDER
<i>Can you really see her as the president?</i> • <i>He could see a great future for her in music.</i> • <i>She's got a new book coming out but I can't see it doing very well.</i>	IMAGINE
<i>As we saw in Chapter 2, the reasons for the war were complex.</i>	FIND OUT INFORMATION
<i>He saw action in Vietnam.</i> • <i>The region has seen some of the fiercest fighting in the war.</i>	EXPERIENCE SOMETHING

Apart from sharing some features with traditional entries in learner's dictionaries (such as the structure of microentries and the use of signposts), the method of lexicographic description of a polysemous word based on its LP presented in this paper also has a number of differences. The main difference is the absence of a list of definitions. The entry begins with a formula for LP of the headword, followed by the three types of meanings described above. The PM which describes the prototypical sense — *perceive with the eyes* — serves as the starting point for the formation of all LSVs. Unlike the PM, which refers to a state, in its metonymic senses the verb *see* denotes actions — *watch, look up, visit, etc.* — which are nonetheless associated with visual perception.

Metaphorical senses are related to the PM through the feature of comparison [AS IF/AS IT WERE]. The metaphorical nature of the actions falling under this definition is generally determined by the nature of the objects of perception — abstract concepts that are inaccessible to visual perception. Therefore, in all meta-

phorical uses, the verb *see* refers to mental processes that the speaker conceptualises as *perceive with the eyes, as it were*.

The features that underlie metonymic and metaphorical transfers are indicated in the entry by means of signposts. These signposts do not serve as definitions, they are intended to help the dictionary user navigate the entry. Supplemented by illustrative examples, they serve as contextual cues that enable the user to infer the contextual meanings of the polysemous headword. They reveal, clarify and actualise a certain aspect of the LP, which may help the learner to form a comprehensive idea of the word's meaning.

A dictionary entry of this type also offers the user a fundamentally different perspective on the nature of polysemy: the existence of a systemic meaning and its realisations in speech. The generalised and universal character of the LP formula, which encompasses the primary, metaphorical, and metonymic meanings, demonstrates the motivation of the LSVs by the LP and shows the semantic relations between them. It also helps conceptualise the distinction between metonymic and metaphorical meanings.

From a theoretical point of view, a lexicographic description of this type clearly demonstrates the real non-mirroring of the correlation between the levels of language and speech. At the system level, the meaning of a polysemous word exists as an invariant structure — the LP — while at the speech level it is rendered through a context-based meaning of a specific LSV. Thus, it can be assumed that the proposed dictionary entry reflects the cognitive mechanisms of semiosis.

In conclusion, it should be added that the proposed dictionary entry is a theoretical model that can serve as a basis for further lexicographic research. The suggested approach to the lexicographic description of polysemous words does not imply a rejection of traditional methods. A dictionary compiled according to the principles discussed here can be used as an additional tool for describing polysemous words, alongside traditional dictionaries and teaching materials.

References

- Ageeva, N.G. 1990. *Tipologiya i mekhanizmy glagol'noy metonimii v sovremennom angliyskom yazyke*. Unpublished Candidate of Sciences Dissertation. Kiev: Kiev State Pedagogical Institute of Foreign Languages.
- Alefirenko, N.F. 2005. *Spornye problemy semantiki*. Moscow: Gnozis.
- American Heritage Dictionary*.
<https://ahdictionary.com/> [10 April 2025]
- Apresyan, Yu. D. 1984. *Lingvisticheskoye obespechenye sistemy frantsuzsko russkogo perevoda ETAP-1*. Moscow: Institute of Foreign Languages of USSR Academy.
- Arkhipov, I.K. 1998. Problemy yazyka i rechi v svete prototipicheskoy semantiki. *Problemy lingvistiki i metodiki prepodavaniya inostrannykh yazykov* 6: 5-22.
- Arkhipov, I.K. 2003. *Chelovetchesky faktor v yazyke*. St. Petersburg: NIYaK.
- Arkhipov, I.K. 2004. Kommunikativny tsejnot i prototipicheskaya semantika. *Izvestiya Rossiyskogo gosudarstvennogo pedagogicheskogo universiteta im. A.I. Gertsena* 4(7): 75-84.

- Arkhipov, I.K., D.N. Novikov and S.A. Pesina.** 2006. Organizatsiya I metod postroyeniya slovarnoy statyi na osnove printsypa proto:picheskoy semantyki. Gumanova, M.K. (Ed.). 2006. *Proceedings of the 5th International Conference, 'Yazyki v sovremennom mire'*: 67-73. Moscow: KDU.
- Arnold, I.V.** 2010. Polisemiya sushchestvitelnogo i leksiko-grammaticheskie razryady. Bukharkin, P.E. (Ed.). 2010. *Semantika. Stilistika. Intertekstualnost*: 5-24. Moscow: Knizhny dom "LIBROKOM".
- Arutyunova, N.D.** 1999. *Yazyk i mir cheloveka*. Moscow: Yazuki russkoy kultury.
- Atkins, B.T.S. and M. Rundell.** 2008. *The Oxford Guide to Practical Lexicography*. Oxford/New York: Oxford University Press.
- Brudny, A.A.** 1971. Znachenie slova i psikhologiya protivopostavleniy. Leontyev, A.A. (Ed.). 1971. *Semanticheskaya struktura slova: psikhologicheskie issledovaniya*: 19-27. Moscow: Nauka.
- Brugman, C. and G. Lakoff.** 1988. Cognitive Topology and Lexical Networks. Small, S., G. Cottrell and M. Tannenhaus (Eds.). 1988. *Lexical Ambiguity Resolution*: 477-507. San Mateo: Morgan Kaufman.
- Budagov, R.A.** 1947. *Slovo i ego znachenie*. Moscow: Nauka.
- Cambridge Dictionary.**
<https://dictionary.cambridge.org/> [10 April 2025]
- Chudinov, A.P.** 2003. *Rossiya v metaforicheskom zerkale: Kognitivnoe issledovanie politicheskoy metafory (1991–2000)*. Yekaterinburg: UGPU.
- Collins COBUILD.**
<http://www.collinslanguage.com/collins-elt-learners-of-english/cobuild#.UI8XqFCST8k>
[8 April 2014].
- Cowie, A.P.** 2007. *English Dictionaries for Foreign Learners: A History*. (A Reprint of the 1999 edition.) Oxford: Oxford University Press.
- Devkin, V.D.** 2000. *Ocherki po leksikografii*. Moscow: Prometey.
- Dretske, F.I.** 1969. *Seeing and Knowing*. London: Routledge & Kegan Paul.
- Dubchinskiy, V.V.** 1994. *Iskusstvo sozdaniya slovarey: Konspekt po leksikografii*. Kharkov: KhPU.
- Guillaume, G.** 1973. *Principes de linguistique théorique de Gustave Guillaume*. Québec/Paris: Librairie C. Klincksieck.
- Halas, A.** 2016. The Application of the Prototype Theory in Lexicographic Practice: A Proposal of a Model for Lexicographic Treatment of Polysemy. *Lexikos* 26: 124-144.
<https://doi.org/10.5788/26-1-1338>
- Hanks, P.** 2000. Do Word Meanings Exist? *Computers and the Humanities* 34(1-2): 205-215.
- Jackson, H.** 2002. *Lexicography: An Introduction*. London/New York: Routledge.
- Jakobson, R.O.** 1985. *Izbrannye raboty*. Moscow: Progress.
- Karasev, A.A.** 2011. Definitii v tolkovykh slovaryakh i leksicheskoe znachenie slova. Fetisov, Anatoly (Ed.). 2011. *Proceedings of Herzen Readings: Modern Languages Interuniversity Academic Conference, St. Petersburg, Russia, May 19–20, 2011*: 188-189. St. Petersburg: Herzen State Pedagogical University of Russia.
- Karasev, A.A.** 2013. Polysemous Words Monolingual Learner's Dictionary Entry (Lexical-Prototypical Approach). *St. Petersburg State Polytechnical University Journal: Humanities and Social Sciences* 3(179): 206-210.
- Karasev, A.A.** 2014a. *Biosemioticheskie printsipy uchebnogo odnoyazychnogo slovaryaya (na materiale sovremennogo angliyskogo yazyka)*. Unpublished Candidate of Sciences Dissertation. St. Petersburg: Herzen State Pedagogical University of Russia.

- Karasev, A.A.** 2014b. Uchebnyy odnoiazychnyy slovar' angliyskogo yazyka v svete biosemiotiki. *Nauchnoe mnenie* 4: 202-209.
- Karasev, A.A.** 2014c. Yavlenie polisemii s tochki zreniya biosemiotiki. *Izvestiya Volgogradskogo gosudarstvennogo pedagogicheskogo universiteta. Filologicheskie nauki* 2(87): 16-20.
- Katsnelson, S.D.** 2011. *Soderzhanie slova, znachenie i oboznacheniiye*. Moscow: Editorial URSS.
- Kurylowicz, J.** 1955. Zametki o znachenii slova. *Voprosy yazykoznaniiya* 3: 73-81.
- Leshcheva, L.M.** 1985. *Svyaz' leksicheskikh znacheniy v semanticheskoy strukture (na materiale prilagatel'nykh sovremennogo angliyskogo yazyka)*. Unpublished Candidate of Sciences Dissertation. Minsk: MGLU.
- Levchina, I.B.** 2003. *Razvitie semanticheskoy struktury sinestezicheskikh prilagatel'nykh*. Unpublished Candidate of Sciences Dissertation. St. Petersburg: Herzen State Pedagogical University of Russia.
- Litvin, F.A.** 1984. *Mnogoznachnost' slova v yazyke i rechi*. Moscow: Vysshaya shkola.
- Lodwig, R.R. and E.F. Barret.** 1967. *The Dictionary and the Language*. New York: Hayden.
- Longman Dictionary of Contemporary English.**
<https://www.ldoceonline.com/> [10 April 2025]
- Maturana, H.R.** 1980. Biology of Cognition. Maturana, H. and F.J. Varela (Eds.). 1980. *Autopoiesis and Cognition: The Realization of the Living*: 2-62. Dordrecht: D. Reidel.
- Mednikova, E.M.** 2010. *Znachenie slova i metody ego issledovaniya*. Moscow: Vysshaya shkola.
- Merriam-Webster Dictionary.**
<https://www.merriam-webster.com/> [10 April 2025]
- Oxford English Dictionary.**
<https://www.oed.com/?tl=true> [10 April 2025]
- Pertsov, N.V.** 2001. *Invarianty v russkom slovoizmenenii*. Moscow: Yazyki russkoy kultury.
- Pesina, S.A.** 1998. *Leksicheskiiy prototip v semanticheskoy strukture slova*. Unpublished Candidate of Sciences Dissertation. St. Petersburg: Herzen State Pedagogical University of Russia.
- Pesina, S.A.** 2005a. *Issledovanie semanticheskoy struktury slova na osnove prototipicheskoy semantiki (na materiale angliyskikh sushchestvitel'nykh)*. Unpublished Doctor of Sciences Dissertation. St. Petersburg: Herzen State Pedagogical University of Russia.
- Pesina, S.A.** 2005b. Metodika opredeleniya sodержatel'nogo yadra mnogoznachnogo sushchestvitel'nogo angliyskogo yazyka. *Izvestiya Rossiyskogo gosudarstvennogo pedagogicheskogo universiteta im. A.I. Gertsena* 5(11): 51-59.
- Pesina, S.A.** 2005c. *Polisemiya v kognitionom aspekte*. St. Petersburg: Herzen State Pedagogical University of Russia.
- Ritchie, L.D.** 2013. *Metaphor*. Cambridge/New York: Cambridge University Press.
- Rivelis, E.I.** 2009. Slovar' v svete kognitivnoy lingvistiki. *Studia linguistica cognitiva* 2: *Nauka o yazyke v izmenyayushcheysya paradigme znaniya*: 304-350.
- Shiryaeva, A.V.** 2008. Prototipicheskaya semantika kak osnovanie sinonimii glagolov. *Izvestiya Rossiyskogo gosudarstvennogo pedagogicheskogo universiteta im. A.I. Gertsena* 23(54): 269-274.
- Shmelev, D.N.** 1973. *Problemy semanticheskogo analiza leksiki (na materiale russkogo yazyka)*. Moscow: Nauka.
- Slonimskaya, I.M.** 1990. Leksikograficheskaya teoriya i leksikograficheskaya pragmatika. Perebeynos, V.I. (Ed.). 1990. *Statisticheskaya leksikografiya i uchebnyy protsess*: 48-79. Kiev: KGPIIYa.
- Smirnitky, A.I.** 1954. K voprosu o slove: Problema tozhdestva slova. *Trudy Instituta yazykoznaniiya AN SSSR* 4: 3-49.

- Soldatov, B.G.** 1997. Semiotika sinonimicheskogo opredeleniya v angliyskoy odnoiazychnoy leksikografii. *Proceedings of Aktual'nye problemy teoreticheskoy i prikladnoy leksikografii the 2nd International Workshop on Lexicography, Ivanovo, Russia, 25–27 September 1997*: 60-69. Ivanovo: Ivanovsky State University.
- The Chambers Dictionary.**
<https://chambers.co.uk/> [10 April 2025]
- Ufimtseva, A.A.** 1962. *Opyt izucheniya leksiki kak sistemy*. Moscow: Nauka.
- Ufimtseva, A.A.** 1984. K voprosu o tak nazyvaemom definitsionnom metode opisaniya leksicheskogo znacheniya slova. Yartseva, V.N. (Ed.). 1984. *Slovo v grammatike i slovare*: 134-142. Moscow: Nauka.
- Ufimtseva, A.A.** 2010. *Leksicheskoe znachenie: printsip semiologicheskogo opisaniya leksiki*. Moscow: LIBROKOM.
- Vinogradov, V.V.** 1947. *Russkiy yazyk. Grammaticheskoe uchenie o slove*. Moscow/Leningrad: Nauka.
- Webster's Online Dictionary.**
<http://www.websters-online-dictionary.org/> [8 April 2014]
- Yakovlyuk, A.N.** 2009. Leksiko-semanticheskij variant kak svyazuyushchee zveno mezhdu mnogoznachnym slovom v yazyke i ego realizatsiyey v rechi. *Bulletin of Chelyabinsk State University* 34(172): 156-162.
- Zalevskaya, A.A.** 2000. *Vvedenie v psikholingvistik*. Moscow: Rossiyskiy gosudarstvennyy gumanitarnyy universitet.
- Zaliznyak, A.A.** 2004. Fenomen mnogoznachnosti i sposoby ego opisaniya. *Voprosy yazykoznaniya* 2: 20-45.
- Zvegintsev, V.A.** 1957. *Semasiologiya*. Moscow: Moscovsky Universitet.

Exemplification of Sensitive Words for People with Disabilities in Monolingual English Learner's Dictionaries

Meimei Li, *Center for Linguistics and Applied Linguistics, Guangdong
University of Foreign Studies, Guangzhou,
China (lmm19854460393@163.com)*
(<https://orcid.org/0009-0009-5930-191X>)

Abstract: This paper investigates how sensitive words related to people with disabilities are exemplified in the "Big Five" monolingual English learner's dictionaries. The findings show that learner's dictionaries tend to be cautious in exemplifying such terms, with notable differences in the number of words exemplified and examples provided, as well as in the inclusion of additional examples. Despite these variations, the "Big Five" consistently offer phrase and sentence examples for sensitive words that serve both decoding and encoding purposes. The analysis reveals that the exemplification of sensitive words is influenced by factors such as part-of-speech, attitude labelling, and word currency. In general, adjectival forms of sensitive words are more likely to be exemplified than their nominal counterparts. Sensitive words that remain in current use tend to have higher exemplification rates than those considered old-fashioned. Neutral or euphemistic expressions are more frequently illustrated with examples than terms that carry negative connotations. This paper argues for broader exemplification of sensitive words, particularly neutral and euphemistic ones, and recommends optimising the presentation and the quality of additional examples.

Keywords: BIG FIVE, EXEMPLIFICATION, SENSITIVE WORDS, PEOPLE WITH DISABILITIES, INCLUSIVE LANGUAGE

Opsomming: Toeligting van sensitiewe woorde vir persone met gestremdhede in eentalige Engelse aanleerderswoordeboeke. Hierdie artikel ondersoek hoe sensitiewe woorde wat verband hou met persone met gestremdhede in die "Groot Vyf" eentalige Engelse aanleerderswoordeboeke toegelig word. Die bevindings toon dat aanleerderswoordeboeke daartoe neig om omsigtig te wees in die toeligting van sodanige terme, met opmerkbare verskille in die aantal woorde wat toegelig word en die voorbeelde wat verskaf word, sowel as in die insluiting van bykomende voorbeelde. Ten spyte van hierdie variasie, bied die "Groot Vyf" konsekwent voorbeelde van frases en sinne vir sensitiewe woorde aan wat 'n dekodeerende sowel as enkoderende doel dien. Die analise toon dat die toeligting van sensitiewe woorde beïnvloed word deur faktore soos die woordsoort, etikettering wat ingesteldheid oordra, en hoe aktueel die woord is. Oor die algemeen word byvoeglike vorms van sensitiewe woorde meer toegelig as hulle naamwoordelike teenhangers. Sensitiewe woorde wat steeds in gebruik is, word meer dikwels toegelig as dié wat as outyds beskou word. Neutrale of eufemistiese leksikale items word meer dikwels met voorbeelde

toegelig as items waaraan negatiewe konnotasies geheg word. Hierdie artikel bepleit breër toeligting van sensitiewe woorde, veral neutrale en eufemistiese woorde, en beveel aan dat die aanbieding en kwaliteit van bykomende voorbeelde geoptimaliseer word.

Sleutelwoorde: GROOT VYF, TOELIGTING, SENSITIEWE WOORDE, PERSONE MET GESTREMDHEDE, INKLUSIEWE TAAL

1. Introduction

While the defining and usage labelling of sensitive words in monolingual English learner's dictionaries (MELDs) has been extensively discussed (e.g. Norri 2000, 2020; Van der Meer 2005; Coffey 2010), relatively little scholarly attention has been paid to their exemplification in MELDs. Two studies on the exemplification policy in learner's dictionaries (Stein 2002; Xu 2008) conclude that attitude-labelled and currency-labelled words are more likely to be exemplified than region or subject-marked words. However, both studies selected their target words at random, which naturally led to the inclusion of different types of words. Given the specific nature of sensitive language, findings based on the exemplification of general vocabulary may not be directly applicable to sensitive terms. It is thus unclear: (1) how learner's dictionaries exemplify (or fail to exemplify) sensitive words; (2) whether learner's dictionaries have a consistent exemplification policy for sensitive words; (3) what factors influence the exemplification of sensitive words in learner's dictionaries; (4) how we can improve the exemplification of sensitive words in learner's dictionaries. This paper, therefore, examines the exemplification of sensitive words with varying degrees of social acceptability in the "Big Five" through an analysis of twenty-one terms used to refer to people with disabilities (*cripple, deaf mute, developmentally disabled, differently abled, disabled, dumb, epileptic, handicapped, hard of hearing, insane, invalid, mentally handicapped, mentally ill, midget, neurodivergent, psychopath, retard, retarded, spastic, spaz, and visually impaired*). These terms were selected because they are frequently cited in previous scholarly research (e.g. Anderson and Fox 1988; Norri 2000, 2020; Allan and Burrige 2006) and institutional language guidelines, such as the WHO's *International Classification of Functioning, Disability and Health* (WHO 2013), demonstrating both their prominence in public discourse and their relevance to socially sensitive issues. They were looked up in the online editions of the *Oxford Advanced Learner's Dictionary of Current English* (OALD), the *Longman Dictionary of Contemporary English* (LDOCE), the *Cambridge Advanced Learner's Dictionary* (CALD), the *Collins COBUILD Advanced Learner's Dictionary* (COBUILD), and the second print edition of the *Macmillan English Dictionary for Advanced Learners* (MEDAL2 2007).¹

This article begins with a comparison of the exemplification of sensitive words for people with disabilities in the "Big Five", which is followed by an examination of factors influencing the exemplification in these dictionaries. The study concludes by presenting several tentative suggestions on how to improve the exemplification of sensitive words in learner's dictionaries.

2. Exemplification of sensitive words for people with disabilities in MELDs: Similarities

In order to investigate the exemplification policies adopted in MELDs, I examined the inclusion of examples for the sensitive words selected for the present study. The findings are presented in tabular form (see Table 1) to better summarise the similarities and differences between the dictionaries studied. A plus sign in the table means that the word is exemplified in the dictionary. A minus sign indicates that no example is provided for the word. A zero means that the word or the relevant disability sense is not included in the dictionary. The numbers placed within brackets indicate the number of examples given. For adjectives that can be used nominally as collectives (e.g. *disabled*, *handicapped*, and *insane*), I counted their nominal and adjectival uses related to disability as two distinct entries.

Since the websites of the four learner's dictionaries do not indicate the source of the additional examples, it remains unclear whether these examples are automatically extracted from corpora or selected by dictionary compilers. This lack of transparency raises concerns about their reliability. As Rundell (2015) observes, such examples are often subject to little or no filtering in terms of quality or appropriacy. In light of these concerns, this part of the comparison focuses solely on the examples included in the core entry. The additional examples will be discussed separately in Section 3.

Table 1: Exemplification of sensitive words for people with disabilities in MELDs

Lexical item	Part-of-speech (POS)	OALD	LDOCE	CALD	COBUILD	MEDAL2
cripple	n.	-	-	-	-	-
deaf mute	n.	-	-	-	-	-
developmentally disabled	adj.	0	0	+ (2)	0	0
differently abled	adj.	0	0	+ (2)	0	0
disabled	adj.	+ (7)	+ (4)	+ (1)	+ (1)	+ (1)
	n.	+ (1)	+ (1)	-	+ (1)	-
dumb	adj.	+ (1)	-	+ (1)	+ (1)	-
epileptic	adj.	+ (1)	+ (1)	+ (2)	+ (1)	+ (1)
	n.	-	-	-	0	-
handicapped	adj.	+ (2)	+ (1)	-	+ (2)	-
	n.	-	-	-	+ (1)	-
hard of hearing	adj.	+ (2)	-	+ (1)	+ (1)	-
	n.	+ (1)	-	0	0	0
insane	adj.	+ (2)	+ (1)	+ (2)	+ (1)	+ (1)
	n.	+ (1)	+ (1)	-	+ (1)	0
invalid	n.	+ (3)	+ (1)	+ (1)	+ (1)	+ (1)
mentally handicapped	adj.	-	-	0	0	-
mentally ill	adj.	0	-	0	0	0
midget	n.	-	-	-	-	-

neurodivergent	adj.	+ (2)	0	+ (2)	-	0
psychopath	n.	-	-	-	+ (1)	-
retard	n.	0	0	-	+ (1)	-
retarded	adj.	-	-	-	-	-
spastic	adj.	-	-	-	-	-
	n.	-	-	-	-	0
spaz	n.	0	0	-	0	0
visually impaired	adj.	0	-	+ (2)	0	0
	n.	0	0	+ (2)	0	0
	Words exemplified	11	7	11	12	4
	Examples provided	23	10	18	13	4

A comparison of the exemplification of the selected words in the "Big Five" reveals some similarities in how sensitive words for people with disabilities are exemplified. These similarities pertain to the number of examples provided, the types of examples used, and the functions those examples serve.

2.1 Number of examples

While the importance of examples in dictionaries has been emphasised by many lexicographers (e.g. Zgusta 1971: 263-268; Landau 2001: 305-308; Atkins and Rundell 2008: 452-462), sensitive terms for people with disabilities, particularly those considered offensive (e.g. *cripple*, *deaf mute*, *midget*, *retarded*, and *spastic*), appear to be exemplified less frequently in learner's dictionaries. The exemplification of the selected words in the "Big Five" reveals an overall lack of examples and unevenness in the coverage of examples.

Table 2 presents the exemplification frequency of the selected words in five MELDs. The data in the table shows a significant lack of examples for disability-related words in learner's dictionaries. While these dictionaries include between 18 and 25 words, the percentage of exemplified words varies widely, with some dictionaries providing very few examples. The highest rate of exemplification is found in COBUILD (63%), followed by OALD (52%), CALD (44%), LDOCE (32%), and MEDAL2 (22%). This means that even in the best case, more than a third of the words surveyed lack illustrative examples.

Table 2: Frequency of exemplification of words for people with disabilities in MELDs

	OALD	LDOCE	CALD	COBUILD	MEDAL2
Words included	21	22	25	19	18
Words exemplified	11	7	11	12	4
Percentage of exemplification	52%	32%	44%	63%	22%

Furthermore, there are significant variations in how specific words are exemplified in different dictionaries. *Disabled* (adj.), *epileptic* (adj.), *insane* (adj.), and

invalid (n.) are the only words consistently illustrated with examples in all five dictionaries. In contrast, *cripple* (n.), *deaf mute* (n.), *midget* (n.), *retarded* (adj.), and *spastic* (adj.) are not exemplified in any of them. The remaining words are treated differently across the five dictionaries.

One possible explanation for the absence of examples for sensitive terms in learner's dictionaries might be the limited corpus evidence available for such words. Many of these terms, particularly those targeting marginalised groups, occur infrequently in general corpora, making it difficult for lexicographers to extract appropriate corpus examples to illustrate their usage. Moreover, slang terms such as *spaz* are typically found in spoken language, which tends to be underrepresented in most corpora compared to written language. This further constrains the availability of suitable examples. However, limited corpus evidence alone may not fully account for the dearth of examples. For instance, while *cripple* (n.) appears 125 times in the *British National Corpus* (BNC), it is not exemplified in any of the "Big Five", whereas *spaz* (n.), which appears only once in the BNC, is exemplified in one of them (CALD). This suggests that the lack of examples for sensitive words may not be solely attributable to corpus frequency, and that other factors may also be at play.

A more compelling reason for the scarcity of examples for sensitive words in learner's dictionaries may be some sensitivity concerns. Lexicographers tend to take a cautious approach to sensitive words in case they might be criticised for normalising offensive language. Hughes (2006: 128) discusses the potential risk of including offensive terms in dictionaries:

Although modern dictionaries are predicated on the assumption that they are *descriptive* — that is, that they should reflect actual usage — there are both lingering and new social pressures about what is appropriate to appear in print. There are also tenacious assumptions that simply by printing an offensive term a dictionary is in some way dignifying the term, relaxing standards, encouraging laxity, or endorsing prejudices.

Dictionaries have to accordingly use various strategies to respond to these pressures, one of the most common being the exclusion of usage examples.

A final possible reason for the limited exemplification of sensitive words might be concern about their potential misuse. Rather than risk reinforcing negative usage or failing to convey the nuanced contexts in which such terms might appear, lexicographers may choose to avoid providing examples altogether, even when corpus evidence is available.

2.2 Types of examples

Learner's dictionaries also show similarities in the types of examples they provide. They all include both phrase examples and sentence examples for words related to people with disabilities. The number of each type of example is shown in Table 3.

Table 3: Number of different types of examples in MELDs

	OALD	LDOCE	CALD	COBUILD	MEDAL2
Phrase examples	7	5	1	3	1
Sentence examples	16	5	17	10	3
Total	23	10	18	13	4

As shown in Table 3, the "Big Five" generally give more sentence examples than phrase examples for the words examined. Phrase examples show typical collocations of the word examined or fixed expressions, such as the examples under *dumb*, *disabled*, and *epileptic* shown in (1), (2), and (3). Sentence examples, on the other hand, place the word in a specific context, as the examples for *differently-abled* and *insane* in (4) and (5) show.

- (1) a young deaf and dumb man (COBUILD)
- (2) a new home for severely disabled people (OALD)
- (3) an epileptic fit (CALD)
- (4) The same opportunities must be provided to differently-abled employees as to the rest of the workforce. (CALD)
- (5) The killer was declared criminally insane. (LDOCE)

Given the particular complexity and sensitivity of sensitive words, it is worth considering that future online learner's dictionaries could benefit from including more sentence examples for such terms. Sentence examples offer valuable contextual cues, such as syntactic patterns, collocational behaviour, and pragmatic implications. For sensitive or potentially offensive terms, understanding the context in which a word is typically used is crucial. By illustrating how a term is used in real-life situations, examples can help learners not only grasp its meaning but also develop the pragmatic competence necessary to avoid inappropriate or harmful usage. Moreover, several dictionary user studies have shown that EFL learners tend to favour full-sentence examples over short segments (e.g. Farina et al. 2019). This preference further supports the recommendation that online learner's dictionaries should consider providing more sentence examples for sensitive words.

2.3 Functions of examples

The functions of dictionary examples have been stressed by many lexicographers (e.g. Fox 1987; Atkins and Rundell 2008). Fox (1987: 137) argues that dictionary examples should "help to reinforce the meaning — not by acting as a reformulation of the definition, but by showing how the word is actually used, in an appropriate context, a typical grammatical structure, and with words that are normally associated with it". Atkins and Rundell (2008: 453-454) summarise the functions of a good dictionary example as including attestation, elucidating the meaning of words, and illustrating contextual features. However, these accounts have been criticised for failing to differentiate examples used for decoding pur-

poses and those for encoding purposes (Humb   2001; Stein 2002; Frankenberg-Garcia 2012, 2014). Thus, Humb   (2001) argues for a clear distinction between decoding and encoding examples.

The "Big Five" provide both decoding and encoding examples for the words examined, especially for neutral terms. Decoding examples are typically designed to offer contextual clues that help learners comprehend word meanings (Humb   2001). The following are examples listed under *invalid* in the "Big Five":

- (6) She had been a delicate child and her parents had treated her as an invalid. (OALD)
- (7) He was found guilty of murdering his invalid father. (OALD)
- (8) Her husband was an invalid and couldn't come to the door to open it. (OALD)
- (9) I resented being treated as an invalid. (LDOCE)
- (10) Is the invalid in bed? (CALD)
- (11) I hate being treated as an invalid. (COBUILD)
- (12) He had been an invalid for many years. (MEDAL2)

Examples (6) and (8) provide relatively clear contexts that help users infer the meaning of *invalid* through contextual cues. In contrast, examples like (7), (9), and (11) either focus on emotionally charged or abstract experiences, making the word's meaning harder to extract. Others, such as (10), lack sufficient context, reducing their effectiveness for modern learners. Ideally, examples provided in dictionaries should be natural, typical, informative, and intelligible (Atkins and Rundell 2008). However, the examples for sensitive words in the "Big Five" vary in clarity and pedagogical usefulness. While some provide meaningful and context-rich usage, others are vague or emotionally complex, potentially hindering comprehension. This variation highlights the importance of careful example selection, especially for potentially offensive words.

In contrast, encoding examples, as pointed out by Humb   (2001), should demonstrate grammatical patterns and collocations of the word. Such examples are mainly given to help facilitate production. Some examples listed under *disabled* are shown below.

- (13) physically/developmentally/intellectually disabled (OALD)
- (14) a new home for severely disabled people (OALD)
- (15) temporarily/permanently disabled (OALD)
- (16) physically/mentally disabled (LDOCE)
- (17) severely disabled (=unable to move most of your body) (MEDAL2)

Examples (13) to (17) are all encoding examples because they demonstrate common collocations of *disabled* with different modifiers (*physically, developmentally, intellectually, severely, temporarily, permanently, and mentally*). These examples do not seem particularly useful in terms of helping users understand what *disabled*

means. The provision of collocational information is aimed at facilitating users' language production.

It is worth noting that some examples for the words examined can deal with both the decoding and encoding needs of users, as the example for *disabled* in (18).

- (18) If you are elderly or physically disabled, massage can be beneficial.
(LDOCE)

The example primarily serves a decoding purpose because it helps users understand the meaning of *disabled* by linking it to a clear health-related context. However, it also has an encoding function in that it presents a common collocate of *disabled*, i.e. *physically*, which could help users produce similar sentences.

To conclude, while users of learner's dictionaries may rarely use sensitive words in their productive language, they still need to understand the meanings and usage of these terms when they encounter them. For offensive and taboo words, a clear usage label may often be sufficient to indicate their sensitive nature, especially given the difficulty of crafting example sentences that effectively convey their negative connotations without reinforcing harmful stereotypes. By contrast, exemplification is particularly important for neutral and euphemistic terms related to disability, as these are more likely to be used productively by learners and their appropriate contexts may not be self-evident. It is therefore essential that learner's dictionaries offer well-considered examples for such terms to support inclusive language use.

3. Exemplification of sensitive words for people with disabilities in MELDs: Main variations

Despite general similarities in the inclusion of illustrative examples in learner's dictionaries, the exemplification of disability-related sensitive terms in the "Big Five" shows notable intra-dictionary and inter-dictionary variations. These differences, to a great extent, reflect underlying editorial attitudes toward sensitive language.

3.1 Variations in the number of exemplified words and included examples

Learner's dictionaries differ from each other in the number of exemplified words for people with disabilities and the number of included examples. As Table 1 shows, COBUILD contains the highest number of exemplified words (12), followed by CALD and OALD, each with 11. LDOCE includes fewer exemplified words (7), while MEDAL2, the only print edition among the five dictionaries, has the fewest, with only four.

The limited exemplification of sensitive terms in MEDAL2 can be partly attributed to its print format, which imposes space constraints that may restrict

the number of examples provided. However, this factor alone does not fully explain the lack of exemplification. As previous research has shown, even the online version of MEDAL contains fewer examples than other MELDs, not only for sensitive terms but also for other terms such as academic words (Rees 2024).² This suggests that the sparse exemplification in MEDAL2 is more likely the result of an overall editorial policy rather than a mere consequence of dictionary format.

The relatively high number of illustrative examples in COBUILD reflects the exemplification policy of the dictionary, which states that "examples would be given for most of the words or senses of words explained in it" (Fox 1987: 137). The policy was proposed when the first edition of COBUILD was being planned and has been followed in later editions.

CALD adopts a more inclusive approach toward emerging vocabulary and euphemisms. It includes and exemplifies terms such as *developmentally disabled* and *differently abled*, whereas none of the other dictionaries include these terms. Additionally, among the three dictionaries that include *neurodivergent*, only CALD and OALD provide examples for the word.

While COBUILD exemplifies the most words, the number of examples provided for each sense is generally limited to one. In contrast, although OALD and CALD exemplify fewer words than COBUILD, they provide a greater number of examples. OALD even gives seven examples for the nominal sense of *disabled*. CALD generally provides two examples for most senses. MEDAL2, on the other hand, includes only one example per sense. Although LDOCE exemplifies seven words, it gives one example per sense except for the adjectival sense of *disabled* (with four examples).

3.2 Variations in the inclusion of additional examples

A notable phenomenon of exemplification in online learner's dictionaries is the inclusion of additional examples within the extra sections of the entry. Among the four online learner's dictionaries, three (OALD, CALD, and LDOCE) provide a section of additional examples. Generally, these dictionaries do not provide information on their websites about how examples are selected. It is therefore unclear to users whether these additional examples are automatically generated from corpora or chosen by lexicographers.

OALD provides a separate **Extra Examples** section in addition to the examples included in the core entry (for an example, see the **Extra Examples** for *disabled* in Figure 1). The **Extra Examples** column is presented following the sense it illustrates.

CALD allows users to choose a display of **More examples** or **Fewer examples** following the examples provided within the core entry. An example is the entry for *disabled* shown in Figure 2. There is only one example nested within the core entry. However, if users want to see more examples, they can click on

the **More examples** link, and more examples for the word will be displayed (see Figure 3). Although CALD does not provide any information on the site about the selection of illustrative examples, Wendalyn Nichols, Publishing Manager of the Cambridge Dictionary, confirms that the examples within the core entry in CALD are all chosen by lexicographers in the process of compiling the entries according to the usual principles for learner's dictionaries. The "more examples" are also chosen by lexicographers from the *Cambridge International Corpus* using Sketch Engine (personal communication, March 28, 2025).

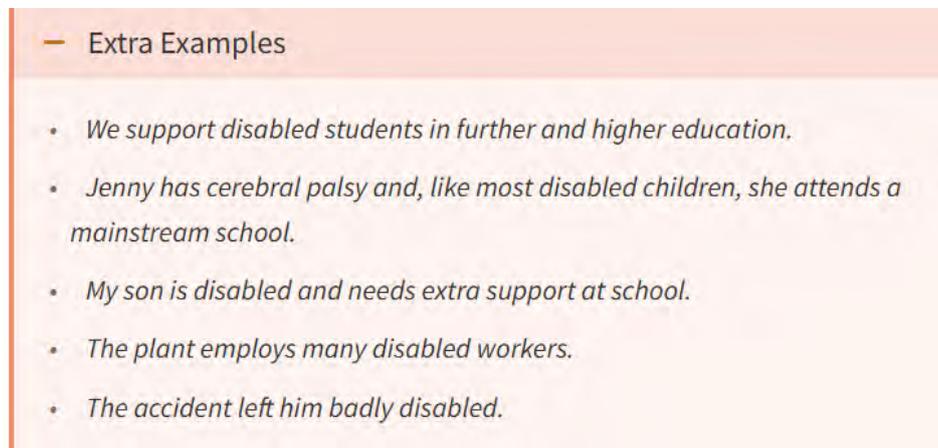


Figure 1: Extra examples for the entry *disabled* in OALD (Accessed 3 June 2025)



Figure 2: The entry for *disabled* in CALD (Accessed 3 June 2025)

– Fewer examples

- *better access for the disabled*
- *This seat is reserved for elderly or disabled passengers.*
- *They have not been very responsive to the needs of disabled customers.*
- *The hotel has good facilities for the disabled.*
- *The company offers holidays for severely disabled people.*

Figure 3: Additional examples for the entry *disabled* in CALD (Accessed 3 June 2025)



Figure 4: Additional examples for the entry *disabled* in LDOCE6 (Accessed 3 June 2025)

LDOCE usually supplements its examples in the core entry with corpus examples, which are explicitly signalled by the boldface subheading **Examples from the Corpus** (for an example see Figure 4). LDOCE provides such additional examples for eleven terms surveyed in this study (*cripple*, *deaf mute*, *disabled*, *dumb*, *epileptic*, *handicapped*, *insane*, *invalid*, *mentally handicapped*, *psychopath*, and *visually impaired*). However, it is worth mentioning that LDOCE's corpus examples contravene principles of good lexicographic exemplification in several ways, leading to the suspicion that they are automatically generated, with little review from lexicographers. These deviations from good lexicographic exemplification will be examined in detail in the following sections.

Admittedly, the inclusion of extra examples in online dictionaries may benefit dictionary users, given the fact that "the average dictionary user appreciates examples in dictionaries" (Ptasznik 2023: 33). However, the presentation of additional examples in learner's dictionaries also raises all sorts of questions. While OALD and CALD provide additional examples under the senses they illustrate,

LDOCE, by contrast, presents the extra corpus examples at the end of the entry in a way that mixes examples of different senses, making it difficult for users to identify which examples illustrate a particular sense. In other words, users may be confused by the mix of examples for different senses. A more user-friendly approach to presenting corpus-derived examples might be to arrange them under the senses they illustrate or to group these examples according to the order of senses. This would enhance clarity and usability, helping learners better understand word meanings and usage in context.

Another problem is that the additional examples extracted from corpora may contain uncommon words, complex structures, idiomatic phrases, or cultural information which are difficult for learners to understand. For example, the following additional examples for *epileptic* and *invalid* are presented on the web page of LDOCE.

- (19) Thus **epileptic** seizures often began with a sustained tachycardia in spite of apnoeic pauses and severe hypoxaemia.
- (20) Ten patients had hypoxaemic events induced by **epileptic** seizures.
- (21) The survey attempted to answer critics who have dismissed international comparisons as **invalid** because of differences in cultural expectations about health care.
- (22) As it turns out, the fears that govern such organizations derive in large part from **invalid** or negative core beliefs.

The highly technical words, such as *tachycardia*, *apnoeic*, and *hypoxaemia*, in the examples for *epileptic* are difficult for foreign language learners — and maybe native speakers too — to understand. Moreover, the complex sentence structures adopted in these corpus examples, especially the example (21), may confuse learners rather than help them understand the meaning of the word. Such examples are largely useless, as Atkins and Rundell (2008: 461) argue that "if the example is incomprehensible, it is of no value".

However, a more serious problem is that some examples do not correspond to the senses they are intended to illustrate. For example, in examples (21) and (22), *invalid* is used in the general sense of "wrong" rather than in a context related to disability, and therefore does not represent a sensitive use. This is a common problem in e-lexicography (Frankenberg-Garcia et al. 2021).

A final deficiency of additional examples is that, without human intervention, they may contain grammatical errors, spelling errors, and other problems, as shown by the following example under *visually impaired* in LDOCE.

- (23) Instructions can also be obained in Braille for the visually impaired.

The word "obained" in this example is undoubtedly a misspelling: it should be "obtained". Such an example is unhelpful to dictionary users. It should be noted that all examples provided in dictionaries, whether selected by lexicographers or generated from a corpus, should serve the needs of dictionary users.

If learner's dictionaries choose to include additional examples, human proof-reading is desirable to ensure that poor-quality examples are eliminated. Admittedly, human intervention comes with financial costs. Given the current challenges facing dictionary publishers, such as increased competition from AI chatbots, declining advertising revenue, and the widespread expectation that dictionaries should be freely accessible, comprehensive manual oversight is often economically unfeasible. Nevertheless, intervention from lexicographers can undoubtedly improve the quality of corpus-derived examples, thereby benefiting foreign language learners.³

4. Exemplification of sensitive words for people with disabilities in MELDs: Main factors

Based on Stein (2002) and Xu (2008), I examined factors that may influence the exemplification of sensitive words for people with disabilities. The analysis indicates that the exemplification of sensitive words in the "Big Five" correlates with the parameters of part-of-speech, attitude labelling, and word currency. This section discusses how the exemplification of selected words varies with respect to these variables.

4.1 Effects of part-of-speech on exemplification

A preliminary analysis of the data in Table 1 reveals that the exemplification rates of sensitive words in the "Big Five" generally vary with the part-of-speech of the word. Adjectival senses of sensitive words are more likely to be exemplified than their nominal counterparts. For example, the adjectival senses of the words such as *epileptic*, *disabled*, and *insane* are consistently exemplified across all five dictionaries, whereas their nominal senses are either not included or not exemplified.

To quantify this exemplification trend, a comparative analysis was conducted (see Table 4). Across the dataset presented in Table 1, there are 14 adjectival terms and 14 nominal terms. The adjectival terms receive a total of 29 instances of exemplification (out of a possible 70), resulting in an average exemplification rate of 41%. In contrast, the nominal terms receive only 16 instances of exemplification (23%). This substantial difference suggests a consistent pattern favouring adjectives in the provision of usage examples.

Table 4: The exemplification rates of adj. and n.

POS	Total lexical items	Total exemplified instances	% Exemplified
adj.	14	29	41%
n.	14	16	23%

To determine whether this difference is statistically significant, a chi-square test

was performed. The result ($\chi^2 = 4.72, p < .05$) confirmed that the association between part-of-speech and exemplification is significant. In other words, the part-of-speech does have a significant effect on the exemplification of sensitive words in learner's dictionaries.

Despite this exemplification pattern, there are still some discrepancies in the exemplification of several words that can be used as both adjectives and nouns across the "Big Five" (e.g. *disabled*, *handicapped*, *hard of hearing*, and *insane*). The adjectival sense of *disabled* is exemplified in all five dictionaries, showing consistency in its treatment as an adjective. The nominal sense, however, is exemplified only in LDOCE and COBUILD. For *handicapped*, the adjectival sense is exemplified in OALD, LDOCE, and COBUILD but left unexemplified in CALD and MEDAL2. The nominal sense of the word is even less frequently exemplified. Only COBUILD provides a phrase example ("... measures to prevent discrimination against the handicapped."). As for *hard of hearing*, the adjectival use is exemplified in OALD, CALD, and COBUILD but unexemplified in LDOCE and MEDAL2. The nominal sense, however, is included in OALD and LDOCE, with illustrative examples provided only in OALD. The other dictionaries do not include the nominal usage at all. When it comes to the exemplification of *insane*, the adjectival sense is consistently exemplified in all five dictionaries. The nominal form, however, is inconsistently treated: it is illustrated by one example in OALD, LDOCE, and COBUILD, respectively, but not in CALD. MEDAL2 does not include this sense. In brief, COBUILD tends to exemplify the nominal uses of sensitive words more frequently than other dictionaries, while MEDAL2 exemplifies them the least.

Nevertheless, it is still safe to conclude that adjectival forms of sensitive words are more frequently exemplified than nominal ones. Such differences in the exemplification are largely influenced by the evolution of disability models and disability language. Nominal forms such as *a cripple*, *an epileptic*, or *a retard* are rooted in the moral model of disability, which views "disability [as] a defect caused by moral lapse or sins" (Olkin 1999: 25). The stigma-based language such as *a cripple* often reflects pity, stigma, or moral judgment (Olkin 2017). Moreover, the expressions like *the disabled* are favoured by the medical model of disability, which views disability as an inherent medical problem of the individual that must be "cured or eliminated" (Siebers 2008: 3). Using nominal forms with the definite article to group people based on certain characteristics like disability has been criticised for reducing individuals to a single attribute and failing to acknowledge their personhood (Halmari 2011; Dunn and Andrews 2015). The affirmative model represents a more recent and influential approach to understanding disability. It emphasises the positive identity and experiences of disabled individuals. Proponents of the affirmative model advocated for the identity-first language (e.g., *blind person* and *disabled person*), acknowledging disability as an integral part of identity, similar to race or gender. To recap, developments in both disability models and disability language offer insight into why adjectival forms of sensitive terms are more frequently exemplified than their nominal counterparts in learner's dictionaries.

4.2 Effects of attitude labelling on exemplification

In order to investigate the effects of attitude labelling on the exemplification of sensitive words for people with disabilities in the "Big Five", I examined the exemplification of the words labelled as having negative connotations by means of usage labels (e.g. "disapproval", "disapproving", "impolite", "not polite", "offensive", "rude", and "taboo") and other devices like definitions and usage notes. The results are tabulated in Table 5. In the Table, "L." means labelling, whereas "Ex" indicates exemplification. In the "L." column, a plus sign means that the word's offensiveness is marked by an attitude label in the dictionary. A minus sign shows that no label is provided for the word. A minus sign placed within brackets indicates that the word's offensiveness is marked by other devices such as definitions or usage notes. A zero means that the word or the relevant disability sense is not included in the dictionary.

Table 5: Effects of attitude labelling on exemplification

Lexical item	POS	OALD		LDOCE		CALD		COBUILD		MEDAL2	
		L.	Ex.	L.	Ex.	L.	Ex.	L.	Ex.	L.	Ex.
cripple	n.	+	-	(-)	-	+	-	+	-	+	-
deaf mute	n.	+	-	+	-	-	-	+	-	(-)	-
developmentally disabled	adj.	0	0	0	0	-	+	0	0	0	0
differently abled	adj.	0	0	0	0	-	+	0	0	0	0
disabled	adj.	-	+	-	+	-	+	-	+	-	+
	n.	-	+	-	+	+	-	(-)	+	(-)	-
dumb	adj.	+	+	(-)	-	-	+	+	+	(-)	-
epileptic	adj.	-	+	-	+	-	+	-	+	-	+
	n.	-	-	-	-	-	-	0	0	-	-
handicapped	adj.	+	+	(-)	+	+	-	+	+	(-)	-
	n.	+	-	(-)	-	+	-	+	+	(-)	-
hard of hearing	adj.	-	+	-	-	-	+	-	+	-	-
	n.	-	+	-	-	0	0	0	0	0	0
insane	adj.	+	+	-	+	+	+	-	+	-	+
	n.	+	+	-	+	-	-	(-)	+	0	0
invalid	n.	-	+	-	+	-	+	-	+	-	+
mentally handicapped	adj.	+	-	-	-	0	0	0	0	-	-
mentally ill	adj.	0	0	-	-	0	0	0	0	0	0
midget	n.	+	-	+	-	+	-	+	-	+	-
neurodivergent	adj.	-	+	0	0	-	+	-	-	0	0
psychopath	n.	-	-	-	-	-	-	-	+	-	-
retard	n.	0	0	0	0	+	-	+	+	+	-
retarded	adj.	+	-	(-)	-	+	-	+	-	(-)	-
spastic	adj.	+	-	-	-	+	-	-	-	+	-
	n.	+	-	-	-	+	-	+	-	0	0
spaz	n.	0	0	0	0	+	-	0	0	0	0
visually impaired	adj.	0	0	-	-	-	+	0	0	0	0
	n.	0	0	0	0	(-)	+	0	0	0	0

As Table 5 indicates, the "Big Five" show discrepancies in their exemplification of attitude-labelled sensitive words. Among them, CALD, LDOCE, and MEDAL take a more consistent approach by generally refraining from providing examples for words labelled as having negative connotations. COBUILD exemplifies more words in general, with only six of the words lacking examples: *cripple*, *deaf mute*, *midget*, *neurodivergent*, *retarded*, and *spastic*. In contrast, OALD is the most inconsistent, which makes it difficult to identify a clear principle behind its exemplification choices. Some words labelled as having negative connotations, like *dumb* (in its "unable to speak" sense), are accompanied by illustrative examples, while others (e.g. *cripple* and *deaf mute*) are not.

Notwithstanding these differences, the "Big Five" tend to adopt a cautious and selective exemplification of attitude-labelled words. To illustrate this, I tabulated the exemplification rates of attitude-labelled words and unlabelled words in Table 6.

Table 6: Exemplification rates of attitude-labelled words and unlabelled words

	Exemplified	Not exemplified	Total	% Exemplified
Labelled words	13	39	52	25%
Unlabelled words	32	21	53	60%
Total	45	60	105	43%

Based on 105 valid instances where the relevant disability-related sense of a word is included, attitude-labelled words have a significantly lower exemplification rate (25%) than unlabelled words (60%). A chi-square test of independence confirmed that this difference is statistically significant ($\chi^2(1, N = 105) = 13.41, p < .001$). This suggests that the attitude-labelled words are less likely to be exemplified in the "Big Five" than unlabelled words.

4.3 Effects of word currency on exemplification

Following Stein (2002) and Xu (2008), I also examined the effects of word currency on the exemplification of sensitive words. I surveyed the example allocation of sensitive words labelled as 'old-fashioned' (Lo.). Table 7 presents data on whether each sensitive word is marked as 'old-fashioned' (Lo=+) and whether it is accompanied by an example sentence (Ex=+). The results of the exemplification rates of currency-labelled words and unlabelled words are listed in Table 8.

It can be seen in Table 7 that the "Big Five" tend to avoid exemplifying highly offensive terms that are now outdated (e.g. *cripple*, *deaf mute*, *dumb*, *midget*, *mentally handicapped*, *retarded*, and *spastic*). By contrast, sensitive words that remain in medical, legal, or formal use (e.g. *invalid* and *insane*) are generally exemplified in MELDs. Among the 105 dictionary entries included in the analysis, 38 are

labelled as old-fashioned, while 67 are unlabelled. Of the labelled entries, only 10 (26%) are accompanied by examples, compared to 35 (52%) of the unlabelled entries. These descriptive results suggest that sensitive words labelled as old-fashioned are less likely to be exemplified than unlabelled words.

To determine whether this observed difference was statistically significant, a chi-square test of independence was conducted. The results revealed a statistically significant association between currency labelling and exemplification rates, $\chi^2(1, N = 105) = 5.64, p < .05$. This indicates that unlabelled words, i.e. more current in usage, are significantly more likely to be provided with examples in learner's dictionaries than words explicitly labelled as old-fashioned.

Table 7: Effects of currency labelling on exemplification

Lexical item	POS	OALD		LDOCE		CALD		COBUILD		MEDAL2	
		Lo.	Ex.	Lo.	Ex.	Lo.	Ex.	L0.	Ex.	L.	Ex.
cripple	n.	+	-	+	-	+	-	-	-	-	-
deaf mute	n.	+	-	+	-	+	-	+	-	+	-
developmentally disabled	adj.	0	0	0	0	-	+	0	0	0	0
differently abled	adj.	0	0	0	0	-	+	0	0	0	0
disabled	adj.	-	+	-	+	-	+	-	+	-	+
	n.	+	+	-	+	-	-	-	+	-	-
dumb	adj.	+	+	+	-	+	+	-	+	+	-
epileptic	adj.	-	+	-	+	-	+	-	+	-	+
	n.	-	-	-	-	-	-	0	0	-	-
handicapped	adj.	+	+	+	+	+	-	-	+	+	-
	n.	+	-	+	-	+	-	-	+	-	-
hard of hearing	adj.	-	+	-	-	-	+	-	+	-	-
	n.	-	+	-	-	0	0	0	0	0	0
insane	adj.	+	+	-	+	+	+	-	+	+	+
	n.	+	+	-	+	+	-	-	+	0	0
invalid	n.	-	+	-	+	+	+	-	+	-	+
mentally handicapped	adj.	+	-	+	-	0	0	0	0	-	-
mentally ill	adj.	0	0	-	-	0	0	0	0	0	0
midget	n.	-	-	-	-	-	-	-	-	-	-
neurodivergent	adj.	-	+	0	0	-	+	-	-	0	0
psychopath	n.	-	-	-	-	-	-	-	+	-	-
retard	n.	0	0	0	0	-	-	-	+	-	-
retarded	adj.	+	-	+	-	+	-	+	-	+	-
spastic	adj.	+	-	+	-	+	-	-	-	-	-
	n.	+	-	-	-	+	-	-	-	0	0
spaz	n.	0	0	0	0	-	-	0	0	0	0
visually	adj.	0	0	-	-	-	+	0	0	0	0
impaired	n.	0	0	0	0	-	+	0	0	0	0

Table 8: Exemplification rates of currency-labelled words and unlabelled words

	Exemplified	Not Exemplified	Total	% Exemplified
Labelled words	10	28	38	26%
Unlabelled words	35	32	67	52%
Total	45	60	105	43%

5. Exemplification of sensitive words for people with disabilities in online MELDs: Some suggestions

Given the importance of examples in learner's dictionaries and the particular nature of sensitive words, improving their exemplification in MELDs is essential. This section thus provides some suggestions on the exemplification of sensitive words in online learner's dictionaries.

5.1 Increasing the exemplification of sensitive words

It may be difficult for learners to grasp the pragmatics of sensitive words without the help of examples. They may run the risk of using these words inappropriately, thereby causing unintended offence to others. It is thus reasonable to argue that learner's dictionaries could exemplify a greater number of sensitive words and increase the number of examples per sense. Traditional print dictionaries, due to space constraints, often grapple with the question of what types of vocabulary should be prioritised in exemplification (Xu 2008). Sensitive words, especially offensive words or taboo words, are seldom exemplified in print dictionaries (for example, in MEDAL2). However, in an era of e-lexicography, storage space is no longer a major problem for lexicographers. Lexicographers should now focus on what types of examples should be provided for different words and how many examples are needed to benefit dictionary users.

Current learner's dictionaries usually provide one or two examples per sense for sensitive words. However, numerous dictionary user studies (e.g. Summers 1988; Laufer 1993; Nesi 1996; Al-Ajmi 2008; Frankenberg-Garcia 2015) have demonstrated the effects of dictionary examples in language learning. Some studies (Frankenberg-Garcia 2014; Ptasznik 2023) show that more examples may benefit dictionary users. Zgusta (1971: 264) states outright that "exemplification is always useful". I would thus argue for an increase in the number of examples per sense for sensitive words (especially neutral and euphemistic terms) in learner's dictionaries.

While storage space is no longer a major limitation in online dictionaries, presentation space remains restricted (Lew in press).⁴ Providing too many examples within a single entry may lead to information overload, potentially hindering rather than helping users (Lew in press; Gouws and Tarp 2017). Therefore, examples for sensitive words in online learner's dictionaries should be selectively

and purposefully chosen to enhance understanding and avoid the dangers of information overload.

5.2 Improving the quality of additional examples

Considering the deficiencies in the additional examples in online learner's dictionaries, there is a clear need to improve the quality of such examples. One efficient solution could be the use of example extraction tools to identify and select good corpus examples. One of the most widely used tools is GDEX (Good Dictionary Examples; Kilgarriff et al. 2008), which was originally developed and used for the *Macmillan English Dictionary* in 2007 (Kilgarriff et al. 2008; Kosem et al. 2019). Significantly, GDEX can be used to filter out examples that do not meet specific standards. For example, it can control sentence length and word frequencies (Kilgarriff et al. 2008). Therefore, learner's dictionaries could use such example extraction tools to select high-quality corpus examples for sensitive words. For instance, examples could be limited to sentences using the defining vocabulary, which would help avoid the use of uncommon words in examples. Furthermore, human proofreading should be involved to check the quality of corpus-derived examples. This would mostly eliminate typos like *obained* mentioned above. To sum up, a combination of human intervention and the use of example extraction tools would largely prevent the problems associated with additional examples in current learner's dictionaries and provide real benefit to users.

5.3 Optimising the presentation of additional examples

As discussed earlier, online learner's dictionaries like LDOCE place additional examples at the end of the entry, where examples of different senses are mixed together. This undifferentiated presentation makes it difficult for users, particularly EFL learners, to identify the particular senses different examples exemplify. Yamada (2013: 223) argues that the improvements of information provision in online dictionaries should focus more on "the order, arrangement, hierarchy, and presentation with differentiation and options". It is therefore argued that the presentation of additional examples in learner's dictionaries should be optimised. One possible approach, if the dictionary prefers to group extra examples together as in LDOCE, is to arrange them in the order of the senses they exemplify. Alternatively, as seen in OALD and CALD, the additional examples could be placed immediately after the sense they are intended to illustrate. Furthermore, in the digital environment, dictionaries can implement interactive features that allow users to filter or highlight examples by sense. For example, users could click on a specific sense in the definition area to view only the examples linked to that meaning.

6. Conclusion

This paper has examined the exemplification of sensitive words for people with disabilities in the "Big Five". While online learner's dictionaries tend to provide more examples overall, they remain conservative in exemplifying sensitive medical words. The dictionaries vary in the number of terms exemplified and examples provided, and in the inclusion of additional examples. Despite these differences, the "Big Five" display notable similarities in the types and functions of examples for the words surveyed: all offer phrase and sentence examples that support both the encoding and decoding needs of users. Furthermore, the findings indicate that the exemplification of sensitive words related to disability is largely influenced by three key factors: part-of-speech, attitude labelling, and word currency. First, with regard to part-of-speech, adjectival senses of sensitive words are more likely to be exemplified than their nominal counterparts. Second, sensitive words that remain in current use tend to have higher exemplification rates than those considered old-fashioned. Finally, words that carry negative connotations, typically marked by attitude labels, are less likely to be exemplified than neutral or euphemistic expressions. After a critical examination of the exemplification policies adopted by the "Big Five", this study argues for improving the quality and optimising the presentation of additional examples in online learner's dictionaries. More importantly, it recommends more frequent exemplification of sensitive words, especially neutral and euphemistic terms. As for offensive words, whether they should be exemplified may require further user research.

This study offers a preliminary exploration of the exemplification of sensitive medical words in dictionaries. There are numerous avenues for further research. A survey of users' interpretations of the examples provided for sensitive words in dictionaries would be particularly valuable. Such studies would enable lexicographers to better understand how different users, particularly users with disabilities, interpret and evaluate the lexicographic treatment of sensitive medical words. Besides user research, an investigation into the representation of people with disabilities through sentence examples could also be pursued in future studies. There is also a need for research that could have a greater concern for the rights and experiences of disabled people in lexicography. Such research would contribute to our understanding of disability in lexicographic discourse and promote inclusive language.

Acknowledgements

The author would like to thank Prof. Elsabé Taljard, and three anonymous adjudicators for their insightful comments and suggestions. Special thanks are due to Prof. M. Lynne Murphy of the University of Sussex for her constructive feedback on earlier drafts of this article. This study was inspired by academic discussions with her.

Endnotes

1. The website of the *Macmillan English Dictionary for Advanced Learners* was shut down on 30 June 2023.
2. I am grateful to one anonymous adjudicator for reminding me that MEDAL has fewer examples for words in general, not just sensitive words.
3. Thanks are due to one anonymous adjudicator who called my attention to this point.
4. Lew (in press) explains that "*storage space* refers to the capacity to hold the total content of the dictionary, whereas *presentation space* refers to how much can be presented (displayed, visualized) at a given time to the dictionary user."

References

A. Dictionaries

- CALD.** *Cambridge Advanced Learner's Dictionary*.
<https://dictionary.cambridge.org> [3 June 2025]
- COBUILD.** *Collins COBUILD Advanced Learner's Dictionary*.
<http://www.collinsdictionary.com> [3 June 2025]
- LDOCE.** *Longman Dictionary of Contemporary English*.
<https://www.ldoceonline.com> [3 June 2025]
- MEDAL2.** 2007. *Macmillan English Dictionary for Advanced Learners*. Second edition. Oxford: Macmillan Education.
- OALD.** *Oxford Advanced Learner's Dictionary of Current English*.
<https://www.oxfordlearnersdictionaries.com> [3 June 2025]

B. Other literature

- Al-Ajmi, H.** 2008. The Effectiveness of Dictionary Examples in Decoding: The Case of Kuwaiti Learners of English. *Lexikos* 18: 15-26.
- Allan, K. and K. Burridge.** 2006. *Forbidden Words: Taboo and the Censoring of Language*. Cambridge: Cambridge University Press.
- Anderson, R.C. and R.A. Fox.** 1988. Health Education and the Developmentally Disabled. *Journal of the Multihandicapped Person* 1(4): 327-333.
- Atkins, B.T.S. and M. Rundell.** 2008. *The Oxford Guide to Practical Lexicography*. Oxford/New York: Oxford University Press.
- BNC.** *British National Corpus*. <http://www.natcorp.ox.ac.uk/> [19 June 2025].
- Coffey, S.** 2010. 'Offensive Items', and Less Offensive Alternatives, in English Monolingual Learners' Dictionaries. Dykstra, A. and T. Schoonheim (Eds.). 2010. *Proceedings of the XIV EURALEX International Congress, Leeuwarden, 6–10 July 2010*: 1270-1281. Leeuwarden/Ljouwert: Fryske Akademy — Afûk.
- Dunn, D.S. and E.E. Andrews.** 2015. Person-first and Identity-first Language: Developing Psychologists' Cultural Competence Using Disability Language. *American Psychologist* 70(3): 255-264.

-
- Farina, D.M.T.Cr., M. Vrbinc and A. Vrbinc.** 2019. Problems in Online Dictionary Use for Advanced Slovenian Learners of English. *International Journal of Lexicography* 32(4): 458-479.
- Fox, G.** 1987. The Case for Examples. Sinclair, J.M. (Ed.). 1987. *Looking Up: An Account of the COBUILD Project in Lexical Computing*: 137-149. London: HarperCollins.
- Frankenberg-Garcia, A.** 2012. Learners' Use of Corpus Examples. *International Journal of Lexicography* 25(3): 273-296.
- Frankenberg-Garcia, A.** 2014. The Use of Corpus Examples for Language Comprehension and Production. *ReCALL* 26(2): 128-146.
- Frankenberg-Garcia, A.** 2015. Dictionaries and Encoding Examples to Support Language Production. *International Journal of Lexicography* 28(4): 490-512.
- Frankenberg-Garcia, A., G.P. Rees and R. Lew.** 2021. Slipping Through the Cracks in e-Lexicography. *International Journal of Lexicography* 34(2): 206-234.
- Gouws, R.H. and S. Tarp.** 2017. Information Overload and Data Overload in Lexicography. *International Journal of Lexicography* 30(4): 389-415.
- Halmari, H.** 2011. Political Correctness, Euphemism, and Language Change: The Case of 'People First'. *Journal of Pragmatics* 43(3): 828-840.
- Hughes, G.** 2006. *An Encyclopedia of Swearing: The Social History of Oaths, Profanity, Foul Language, and Ethnic Slurs in the English-speaking World*. New York: M.E. Sharpe.
- Humbly, P.** 2001. *Dictionaries and Language Learners*. Frankfurt am Main: Haag und Herchen.
- Kilgariff, A., M. Husák, K. McAdam, M. Rundell and P. Rychlý.** 2008. GDEX: Automatically Finding Good Dictionary Examples in a Corpus. Bernal, E. and J. DeCesaris (Eds.). 2008. *Proceedings of the 13th EURALEX International Congress, Barcelona, 15-19 July 2008*: 425-432. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.
- Kosem, I., K. Koppel, T.Z. Kuhn, J. Michelfeit and C. Tiberius.** 2019. Identification and Automatic Extraction of Good Dictionary Examples: The Case(s) of GDEX. *International Journal of Lexicography* 32(2): 119-137.
- Landau, S.I.** 2001. *Dictionaries: The Art and Craft of Lexicography*. Second edition. Cambridge: Cambridge University Press.
- Laufer, B.** 1993. The Effect of Dictionary Definitions and Examples on the Use and Comprehension of New L2 words. *Cahiers de Lexicologie* 63(2): 131-142.
- Lew, R.** in press. Space Restrictions in Paper and Electronic Dictionaries and Their Implications for the Design of Production Dictionaries.
https://wa.amu.edu.pl/~rlew/pub/Lew_space_restrictions_in_paper_and_electronic_dictionaries.pdf [19 June 2025]
- Nesi, H.** 1996. The Role of Illustrative Examples in Productive Dictionary Use. *Dictionaries: Journal of the Dictionary Society of North America* 17: 198-206.
- Norri, J.** 2000. Labelling of Derogatory Words in Some British and American Dictionaries. *International Journal of Lexicography* 13(2): 71-106.
- Norri, J.** 2020. Treatment of Words for Illness and Disability in Monolingual English Dictionaries. *International Journal of Lexicography* 33(3): 227-250.
- Olkin, R.** 1999. *What Psychotherapists Should Know about Disability*. New York: Guilford Press.
- Olkin, R.** 2017. *Disability-affirmative Therapy: A Case Formulation Template for Clients with Disabilities*. New York: Oxford University Press.
- Ptasznik, B.** 2023. More Examples May Benefit Dictionary Users. *International Journal of Lexicography* 36(1): 29-55.

- Rees, G.P.** 2024. Academic Word Families in Online English Dictionaries. *Lexikos* 34(1): 437-468.
- Rundell, M.** 2015. From Print to Digital: Implications for Dictionary Policy and Lexicographic Conventions. *Lexikos* 25(1): 301-322.
- Siebers, T.** 2008. *Disability Theory*. Ann Arbor, MI: University of Michigan Press.
- Stein, G.** 2002. *Better Words: Evaluating EFL Dictionaries*. Exeter: Exeter University Press.
- Summers, D.** 1988. The Role of Dictionaries in Language Learning. Carter, R. and M. McCarthy (Eds.). 1988. *Vocabulary and Language Teaching*: 111-125. London: Longman.
- Van der Meer, G.** 2005. On Taboo and Slang in English Pedagogical Lexicography. So You've Been Looking for Them, Have You, Madam? (Dr Johnson). Gottlieb, H., J.E. Mogensen and A. Zettersten (Eds.). 2005. *Symposium on Lexicography XI: Proceedings of the Eleventh International Symposium on Lexicography May 2-4, 2002 at the University of Copenhagen*: 381-390. Berlin/Boston: Max Niemeyer.
- World Health Organization (WHO).** 2013. *How to Use the ICF: A Practical Manual for Using the International Classification of Functioning, Disability and Health (ICF)*. Geneva: WHO.
- Xu, H.** 2008. Exemplification Policy in English Learners' Dictionaries. *International Journal of Lexicography* 21(4): 395-417.
- Yamada, S.** 2013. A Test of the Proposed Framework for Reviewing Online Dictionaries: M-W.com, Dictionary.com, Macmillandictionary.com, Dictionary.cambridge.org, and Oxforddictionaries.com (sic). *Dictionaries: Journal of the Dictionary Society of North America* 34: 211-224.
- Zgusta, L.** 1971. *Manual of Lexicography*. Berlin/New York: De Gruyter Mouton.

Lexicography in Action: The Traversal from Coinage and Iconicity to Iconisation

Dorra Moalla Masmoudi, *LAD Laboratory, Faculty of Arts and
Humanities, University of Sfax, Sfax, Tunisia*
(dorra.moala@flshs.usf.tn)
(<https://orcid.org/0000-0001-8055-2400>)

Abstract: This paper seeks to connect Lexicography and Critical Discourse Analysis (CDA) by highlighting that the critical potential of Lexicography lies in the coinage of dictionary entries, which can function as *semiotic acts* that aim to rally the readers around the sign-maker's world view. To illustrate this, GOT ISRAELED, a new entry in the crowdsourced *Urban Dictionary* added on 21 October 2023, is taken as an example. This entry is defined as the act of allowing someone to share something with you but this person claims the thing as his/her own and expels you. The entry has gone viral on social media and induced divergent attitudes. Its wide reach is due to its *iconic* nature and coinage during the war on Gaza. The entry has also been heavily resemiotised inside *Urban Dictionary* and its critical potential has gradually increased clearly recontextualizing the colonization of Palestine and the issue of Jewish settlements. 36 entries added between 21 October 2023 and 10 July 2024 are analyzed at three levels: *coinage*, *iconicity* and *iconisation*. The analysis draws upon Semiotic and Social Semiotic frameworks, which include DeSaussure's (1916) *sign system*, the Social Semiotic principle of *motivated sign* (Kress 1993), and Peirce's (1931) principles of *hypo-iconicity*. An SFL-based analysis exploiting subsystems within the *ideational* and *interpersonal* metafunctions is employed to explain the sign's 'iconisation' (Halliday and Matthiessen 2014, Kress and Van Leeuwen 1996/2021, Martin and White 2005). The investigation simultaneously highlights the critical potential of the sign and describes its transformation into a *bonding icon* rallying people around the writers' world view (Stenglin 2008, 2012).

Keywords: LEXICOGRAPHY IN ACTION, COINAGE, ICONICITY, ICONISATION, URBAN DICTIONARY, MOTIVATED SIGN

Opsomming: Die leksikografie in aksie: Die oorgang van nuutskepping en ikonisiteit na ikonisering. In hierdie artikel word gepoog om die Leksikografie en Kritiese Diskoersanalise (KDA) met mekaar te verbind. Dit word gedoen deur te beklemtoon dat die kritiese potensiaal van die Leksikografie geleë is in die vaslegging van woordeboekinskrywings wat as *semiotiese handelinge* kan funksioneer. Laasgenoemde is daarop gemik om die lesers te laat saamstem met die sender van die teken se wêreldbeskouing. Ter illustrasie hiervan word "GOT ISRAELED", wat op 21 Oktober 2023 as nuwe inskrywing in die skarebenutte *Urban Dictionary* bygevoeg is, as voorbeeld gebruik. Hierdie inskrywing word gedefinieer as die handeling om iemand toe te laat om iets met jou te deel, maar hierdie persoon eis dan die item as sy/haar eie op en verdryf jou. Die inskrywing het vinnig op sosiale media versprei en uiteenlopende standpunte ontlok. Die breë reikwydte daarvan

is toe te skryf aan die *ikoniese* aard en vaslegging daarvan tydens die oorlog teen Gaza. Die inskrywing is ook in die *Urban Dictionary* grootliks gehersemiotiseer en die kritiese potensiaal daarvan het geleidelik toegeneem om sodoende die kolonisasie van Palestina en die kwessie van Joodse nedersettings duidelik te herkontekstualiseer. 36 inskrywings wat tussen 21 Oktober 2023 en 10 Julie 2024 bygevoeg is, word op drie vlakke geanaliseer: *nuutskeping*, *ikonisiteit* en *ikonisering*. Die analise steun op Semiotiese en Sosiaal-Semiotiese raamwerke, wat DeSaussure (1916) se *tekensisteem*, die Sosiaal-Semiotiese beginsel van *gemotiveerde teken* (Kress 1993), en Peirce (1931) se beginsels van *hipo-ikonisiteit* insluit. 'n SFL-gebaseerde analise wat substelsels binne die *ideasionele* en *interpersoonlike* metafunksies benut, word gebruik om die 'ikonisering' van die teken te verduidelik (Halliday en Matthiessen 2014, Kress en Van Leeuwen 1996/2021, Martin en White 2005). Die ondersoek beklemtoon tegelykertyd die kritiese potensiaal van die teken en beskryf die transformasie daarvan in 'n *saambindende ikoon* wat mense rondom die skrywers se beskouings van die wêreld verenig (Stenglin 2008, 2012).

Sleutelwoorde: DIE LEKSIKOGRAFIE IN AKSIE, NUUTSKEPPING, IKONISITEIT, IKONISERING, *URBAN DICTIONARY*, GEMOTIVEERDE TEKEN

1. Introduction

Lexicography is commonly known as the scholarly field involving compiling dictionaries and editing them (Bergenholtz and Gouws 2012). Mainstream Lexicography prioritizes the objective linguistic descriptions of dictionaries as "objective value-free representations of languages" (Chen 2015: 311) and attributes little attention to the contextualization of dictionary-making. Critical Lexicography, a recent research direction, addresses this gap and showcases that dictionaries are influential ideological texts. Critical Lexicography meshes with CDA, and views language use as a form of social practice conditioned by the context of use and the dominant ideologies (Fairclough 1995). Dictionaries are, therefore, ideological sites that can be understood in terms of the context of use and power dynamics (Bertulessi 2023, Chen 2015, 2019, Fuertes-Olivera and Tarp 2022, Hornscheidt 2008, Pettini 2023). Chen (2015: 362) concentrates on the link between dictionary structures and entries and the external factors of language use stating that "lexicography and its products, dictionaries, are never value-free, apolitical or asocial. Instead, they are subject to ideology, power and politics".

Dictionary writing as "a literate writing activity" is situated within a process of social recontextualization and meaning relocation and "is inevitably bound up with subjectivity, re-representation, objectivity and ideology" (Chen 2015: 312). This entails that dictionary-making is a floating domain that shifts from context to context, practice to another and from sign-maker to sign-maker. According to Linell (1998), recontextualization in Lexicography is concretely realized across three dimensions: intra-textual recontextualization which happens within the same dictionary, intertextual recontextualization which concerns the shifts in meaning-making of the same types of texts situated within different contexts (for example two dictionaries defining the same entry) and interdiscursive recontextualization which occurs across different media and genres (for instance, a dic-

tionary entry and an argumentative text dealing with a dictionary entry). Intra-textual recontextualization is the focus of this paper, a recontextualization afforded by the participatory nature of online crowdsourced dictionaries.

The current literature on Critical Lexicography is centered upon deciphering the ideology related to gender, ethnicity and politics. By means of examining ideology-loaded entries, research studies point out that dictionaries recontextualize ethnic and gender bias in societies (Chen 2015, Fuertes-Olivera and Tarp 2022). These studies have emphasized that ideology is articulated through the selective techniques of the dictionary-maker, which include deleting, replacing, re-signifying ... etc. (Chen 2015). With regards to political ideology, studies have demonstrated that projects of dictionary compilation expose the political views of the dominant political groups (Bertulesi 2023) and perpetuate ideologies of racism and colonialism (Hornscheidt 2008).

The study at hand aligns with Critical Lexicography through showing that ideology stands as a mediating force for the coinage of new signs or the *mutability* of existing ones. The existing literature on Critical Lexicography has adopted a critical stance towards the work of ideology in dictionary-making and has argued that the role of scholars engaging with (Critical) Lexicography research is to uncover the opaque techniques of power dynamics (Chen 2019, Kachru 1995). Though this study acknowledges that Lexicography is a site of ideology encoding, it maintains that dictionaries can also be sites for *semiotic acts*, giving voice to the marginalized and oppressed groups. Chen (2019) articulates the dual work of dictionaries as domains of power and resistance. Within the latter, regional dictionaries, and dictionaries for minority groups can disrupt the hegemony of dominant dictionaries. Resistance Lexicography can be a promising direction of research within Critical Lexicography. To the best of my knowledge, there is a scarcity of substantive research on uncovering the dynamics of meaning-making in *lexicography in action*. This paper aims to address this gap. To do so, it offers to analyze 36 new entries of the lexical item GOT ISRAELED, a new coinage in the online *Urban Dictionary* that first appeared on 21 October 2023 and was widely resemiotised within the dictionary and in different generic and mediated forms outside it. This paper argues that these entries are semiotic acts introducing, in an escalating way, the agency and the subjective views of the compilers, who aim to turn the new entry into a semiotic force that recontextualizes what they consider as an unrightful colonization of Palestine. Within a period of two weeks, the entry had proliferated and shifted, in an incremental way, from an almost apolitical sign to a politicized one overtly recontextualizing the Palestinian cause through pointing to hot political issues, like the Jewish settlements and the religious grounding of the colonization of Palestine. The critical potential of these entries derives from the *iconic* nature of the sign. Studying the critical force of iconic signs based on the resemblance between form and meaning remains underexamined in the literature.

Methodologically speaking, the current studies on Critical Lexicography have pointed out that power manifestations in lexicography are mediated through the selective techniques of censoring, altering, replacing and redesigning (Chen 2019).

This study, however, contends that lexicography, as a semiotic act, exploits techniques of attitude construction and expansion. The compiler's role is to use the discourse resources to rally the reader around his/her world view. In the sign under examination, I propose to delve into how the writers of the different dictionary entries exploit the semiotic principles of *motivated sign*, *iconicity* and *iconisation* to produce an effect upon the reader.

To decipher the dynamics of attitude amassing in the dictionary entries under examination, three levels of analysis are adopted:

- The coinage of the sign
- Its iconicity
- Its iconisation

The first level draws upon De Saussure's (1916) principles of the sign system to highlight the innovative relation between *signifier* and *signified* and it explains, through the principle of the im/mutability of the sign system, the social forces triggering its mutability. Along with this, the analysis adopts a Social Semiotic perspective to underscore the motivated relation between signifier and signified in the coinage of the present sign (Kress 1993).

The second level exploits Peirce's (1931) typology of signs to explore the semiotic relation foregrounded in this iconic sign. By using Peirce's principles of *hypo-iconicity*, this analysis underscores how the principles of *direct resemblance* and *metaphoric distortion* in hypo-iconicity build up direct connection with reality and foreground the critical potential of the sign.

The third level shows how the entry has evolved from iconicity to a *bonding icon* (Stenglin 2022: 314). To highlight this, the analysis draws upon the principles of iconisation, an evolving research direction in Social Semiotics (Stenglin 2008; 2012; Martin 2010). The main tenets of iconisation are that signs turn into bonding icons when they rally people around a set of values (Stenglin 2022). This paper posits that the entry under examination has evolved into a bonding icon through amassing attitudes uniting internet users, though in divergent ways, around the sign. This is translated in its wide resemiotisation inside the dictionary and outside it in different mediated forms, its viral status, as well as the substantial positive and negative reactions the sign has triggered. In the process of its shift into a bonding icon, the entry realizes simultaneously an *ideational* role and an *interpersonal* one: it represents the world from the standpoint of the entry writer and rallies people around this world view. To highlight the interface between ideational and interpersonal meanings in the production of this sign, I adopt an analysis drawing upon the ideational and interpersonal metafunctions in SFL (Systemic Functional Linguistics) (Halliday and Matthiessen 2014, Martin and White 2005). A Systemic Linguistic analysis is undertaken building upon sub-systems within the ideational and interpersonal metafunctions, which include the systems of *represented participants*, *process types*, *participant roles* and *metaphoric distortion*, on the one hand and the systems of *interactive participants* and *graduation*, on the other. The different systems will be explained in details in section 6.

This introduction has argued that a critical reading of lexicography should be combined with the linguistic one to grasp the meaning potential of dictionary entries. Such a critical reading is especially relevant in lexical entries with an action force. Based on these premises, the paper is structured as follows: section 2 grounds the study within the research direction of language as a semiotic act, section 3 introduces the discursive and sociopolitical context of the study. It presents the sociopolitical context of the sign coinage, introduces the dictionary under examination and narrows down the focus to the entries and their generic structure. Sections 4, 5 and 6 analyze the entries at the three levels mentioned above: coinage, iconicity and iconisation using tools from semiotics and SFL. Finally, in section 7, a framework for analyzing signs with an action force is introduced.

2. Semiosis in action

This study draws upon the principle of *semiosis as action* developed by Peirce (1907). By semiosis, Peirce (1907: 411), means "an action, or influence, which is, or involves a cooperation of three subjects, such as a sign, its object, and its interpretant". In this triadic relation, the interpretant's role is central and is understood as the thing that the sign triggers for the interpreter, which can be a feeling, an effect or a thought (Strand 2013). Peirce's semiosis constitutes a paradigm shift from DeSaussure's structure of sign, according to which a sign is inherently communicative as it is a pre-defined and fixed connection between a signifier and a signified (DeSaussure 1916). Peirce's semiotics foregrounds the signs as mediators to action. The meaning of signs is not fixed as the interpreter assigns meanings to them through processes of construction, reconstruction and validation (Strand 2013). The principle of interpreter's role in assigning meaning to signs is congruent with the principle of *reader agency* in Social Semiotics, which advocates that reading is a meaning-making action rather than a reception phenomenon (Kress 2010). The principle of reader agency is more evident in this media age, which affords a more participatory role to the reader (Moalla 2023).

The semiotic effect advocated by Peirce cannot be understood without reference to authorial agency. Social Semiotics describes the role of the designer as semiosis in action or semiosis at work, highlighting the agentive power of the designer in meaning-making and his/her interest in the process of design. Kress and Van Leeuwen (2001) equate semiosis in action with changes in representation, which are done through the action of individuals. This action cannot be dissociated from the macro-level changes in the social context. Kress and Van Leeuwen (2001: 36) describe the role of social forces in semiotic action as "all semiotic action is social: that social action changes the actor, the acted on or acted with".

To sum up, this paper suggests semiosis in action as a framework for understanding the dynamics of meaning-making in the coinage and interpretation of dictionary entries. This framework is aligned with Peirce's semiosis as action

along with the principles of CDA and Social Semiotics. It is understood at two levels. First, in line with CDA, signs are mediated forms that simultaneously recontextualize the world and reflect the authorial voices. Therefore, signs "shape how we think and act in the world" (Ledin and Machin 2018: 15). Second, signs are not fixed entities. They induce action that can be translated into a thought, a feeling or an effect. Of special interest here, is how signs trigger feelings and attitudes. This kind of action is aligned with the principle of iconisation in Social Semiotics, according to which signs align and unite people around a set of values, which are induced by the authorial discursal strategies.

3. Context of the study

Chen (2019) maintains that modern lexicography should be more critical through going beyond the text to the context of text production and its power dynamics. He proposes three phases of lexicographic analysis that include social analysis, discourse analysis and text analysis. This paper aligns with this methodology. This section proposes simultaneously to anchor the sign under examination within the sociopolitical context of its coinage and to ground it within the generic tradition of online dictionaries before undertaking a text analysis in the following sections.

The entry under examination was added to *Urban Dictionary*, an online crowdsourced dictionary about new lexis and slang words (Ro 2019). Internet users can add their lexical entries. After approval, the entry appears with the name of the writer and the date of entry. Internet users can upvote or downvote it and share it on social media platforms. From its creation by the computer science student Aaron Peckham as a humorous parody to scholarly dictionaries, *Urban Dictionary* has become very notorious with 65 million visitors every month. It is also widely used by linguists to keep record of slang words entering into the language (Ro 2019).

Urban Dictionary aligns with online dictionaries, which are increasingly appearing as alternative forms of dictionaries along with the traditional printed ones. Meta-lexicography research on online dictionaries has yielded divided views relative to their textual features and content. Advocates of this emerging genre argue that it allows selected searches of entries and rapid access to them (Kirkness 2004). At the content level, e-dictionaries, especially those treating slang words, can give voice to minority groups to counterbalance the dominance of standard language-based dictionaries (Coleman 2004). Pettini (2023) expands on this, suggesting that printed dictionaries, being produced by multinational publishing industries, reflect the interest of the dominant groups. Online dictionaries give voice to the oppressed and marginalized groups. Critical voices about slang online dictionaries emphasize the poor quality, and the inaccuracies of many of them (Liu 2015) and the lack of a conventionalized structure (Gouws 2014). Critics also contend that online dictionaries proliferate discourses of offensiveness and hate (Pettini 2023).

As to its generic structure, the homepage in *Urban Dictionary* includes four components: (1) the heading of the homepage, (2) a horizontal bar for the menu, (3) a search bar and (4) a right-hand side section for the coinage of new words including also small icons for social media sharing. The dictionary exploits the interactive affordance of the internet allowing users to exert different actions like clicking on the search bar, adding an entry, and sending a question (*Urban Dictionary* 2024). With reference to the expression under consideration, the search bar includes two entries, the verb and noun (Israel and Israeled) along with other compound nouns. The noun 'Israel' includes many entries, mostly coined in the years 2000s and more recent versions were added in November 2023. The definitions have clear political orientations and echo divergent views about the state/entity of Israel. On 21 October 2023, for the first time, a new entry in the verbal form, GOT ISRAELED, was added. This entry is defined as the act of allowing someone to share something with you but this person claims the thing as his/her own and expels you. The sign-maker takes the example of sharing a table in a restaurant with a stranger, who ends up expelling you pretending s/he has a meeting. Five entries were coined on the same day (21 October) by different writers with different pseudonyms. Thus, it is difficult to decide which of them is the original entry, especially that the interdiscursive recontextualization of the entry in other media and genres (online magazines and blogs) refer to one entry appearing on the 22nd as the original. For this reason, I will consider all the entries appearing on 21st as the original entries.

Starting from 22 October and during 9 months, the sign has gone viral on social media and has been widely resemiotised inside *Urban Dictionary*. The resemiotisations have shown, in an escalating way, more politicized definitions of the sign. The link of the sign with the colonization of Palestine becomes more direct. The wide reach of this entry and its viral status are mainly due to its iconic nature and its coinage against the backdrop of the Israeli war on Gaza, which may explain the increasing political orientation. The sign is also resemiotised outside *Urban Dictionary* in different mediated forms (blogs, videos, billboards, online journals) by advocates of the Palestinian cause as well as by defenders of Israel.

Grammatically speaking the verbal form of the entry is used in the passive voice 'be Israeled' in the simple, progressive and perfective aspects (11 entries), in the passive voice with the mutative 'get' as in GOT ISRAELED (16 entries) as well as in the active TO ISRAEL (11 entries). GOT ISRAELED is used generically to refer to the verbal realizations of this entry. Starting from May 30th, six different definitions of the verb had been coined, very probably by pro-Israeli writers, to counter the iconisation of the verb but the definition did not achieve the viral status. Since these entries have a different meaning, they are not within the scope of this study.

The generic structure of each entry aligns with the typical structure of dictionary entries outlined by Fuertes-Olivera and Tarp (2022). It includes a headword highlighting the verbal form of the entry although such grammatical information is not explicitly mentioned as in standardized printed and online diction-

aries. The head word is followed by a co-text, which is divided into a definition followed by a usage example in a concrete situation and ends with a sentence example. The generic structure exhibits a few specificities, which are congruent with the communicative function of crowdsourced dictionaries. As illustrated in the sample entry in Figure 1, the top right corner includes social media links to help the user share the entry. Below the sentence example, the pseudo name of the writer and the date of addition are included. At the bottom of the entry, two icons are provided allowing the user to react by upvoting, or downvoting.

Israeled 🐦 f

When you **allow** someone to use a thing you own or sit in a place that is **reserved** for you then they **claim it's** theirs all along

- Someone asked me to share **my table** at the **coffee shop** and then they asked me to leave the table because they have a meeting?!

- looks like you have been **israeled**

by **Lilylina** October 21, 2023

👍 36006 👎 20933 🚩 FLAG

Get the **Israeled** mug.

Figure 1: A Sample of the dictionary entry ISRAELED included in <https://www.urbandictionary.com/define.php?term=Israeled>

4. Coinage: Mutability and motivation

The entry is considered as a sign that partly subscribes to the structure of the sign in DeSaussure's sign system (1916). A sign, in this framework, is an intrinsic, conventionalized, inherently communicative connection between a cluster of sounds, a signifier and a concept, a signified. The sign under examination is a combination of a signifier in the passive voice with a mutative 'got' (cf. Halliday and Matthiessen 2014: 349) and a signified, 'to be cheated' by someone or a group of people, who take something that used to belong to the victim.

GOT ISRAELED is directly associated with a previous sign in the nominal form (Israel), which is also a controversial sign within *Urban Dictionary*. DeSaussure refers to this shift as the mutability of signs that may affect the signifier, and/or signified and/or the relationship between them (DeSaussure 1916). Despite the phonological resemblance between this sign and the previous nominal form, the mutability occurs at the three levels as shown in Figure 2.

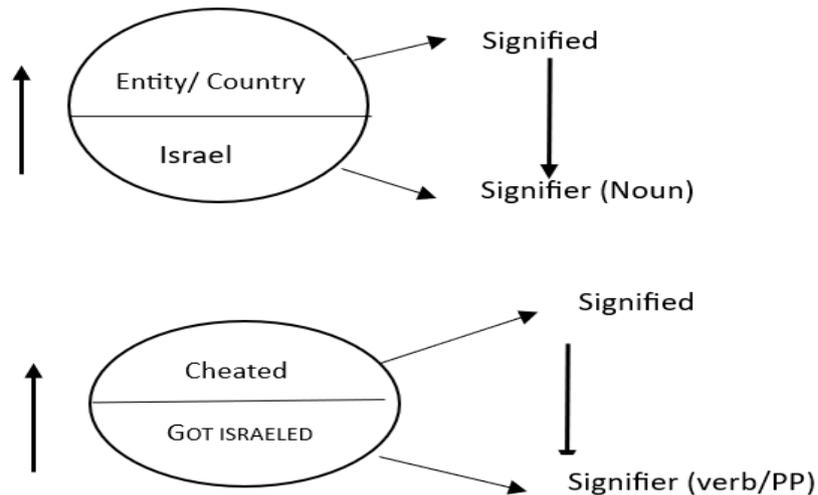


Figure 2: Shift in signifier-signified relationship in GOT ISRAELED

In DeSaussure's framework, language is an inherited immutable system with a fixed relationship between signifier and signified. The individual and linguistic community have little freedom to control or change it. He justifies this fixedness by the fact that "a particular language state is always the product of **historical forces** and these forces explain why the sign is unchangeable, i.e. why it resists any **arbitrary** substitution" (DeSaussure 1916: 72 emphases added). Despite the immutability of signs, DeSaussure argues that language is unable to resist changes at a certain point in time. He posits two factors for sign mutability, which are time and social force. Time allows social force to exert sign mutability but the shift is only possible by the agreement of the community of speakers. His theory provides manifestations of language change (signifier, signified and relationship between them) but its description of time and social force as contributors to sign mutability remains vague and not thoroughly described (Holdcroft 1991).

Many critics attribute DeSaussure's failure to account for language mutability to the fact that he describes the relationship between signifier and signified as arbitrary and unmotivated. On the contrary, Holdcroft (1991: 53) maintains that, beyond onomatopoeia, many signifiers are "appropriate for the signifieds", at different levels of appropriateness, which are divided into categorical and structural ones. Categorical appropriateness refers to the correspondence between grammatical categories and conceptual ones. For instance, nouns denote entities and verbs denote action. Structural appropriateness refers to the link between sentence structure and the structure of thought.

In the Social Semiotic Theory, this appropriateness is conceptualized within the notion of motivated sign (Hodge and Kress 1988, Kress 1993). In this theory, the motivated sign is the result of an interplay between the social context, the material features of signifiers and the individual's contribution to the production of signs (Björkvall and Karlsson 2011). The social/cultural dimension is the superordinate force in meaning-making that dictates social practices and forms of discourse. The making of meaning resides in the intersection between the agreed upon social meaning attributed to signs and the material features of signifiers (for instance, the grammatical category). In this interplay between the social and the structural, a motivated sign is generated. Grammatical and lexical categories and other semiotic modes are culturally established and materially appropriate to express certain meanings. Kress (1993) illustrates this with the example of the circles (signifier) being apt for car wheels (signified). Social practices execute a force in transforming the material features of signifiers into meaning. The individual sign-maker is aware of the fact that signs have a meaning potential dictated by culture/society and the material features of the sign, and exploits this connection to express the desired meanings. The signs produced are "motivated by the producer's interest and the characteristics of the objects" (Kress 1993: 173).

To illustrate this triadic relationship between the social, material and individual in the sign under examination, the sign makers' choice of the root 'Israel' is motivated by the socio-political context of the colonization of Palestine and its appropriateness is accentuated against the backdrop of the war on Gaza. The use of the signifier in the past participle is also materially appropriate to express the signified, the act of being cheated. This sign aligns with Kress's description of motivated sign as transparent in which the reader can easily see why the signifier GOT ISRAELED is used to express the signified. Accordingly, the writers of the entries have drawn upon the choices available in grammar and dictated by context to produce signs. Kress (1993: 174) describes this as a selection process of what to be presented and how to be presented through the use of apt signifiers, which should have "criterial characteristics in respect to the object to be presented, that is, it becomes, the signified". The selection process also reflects the sign maker's relation to an object or event at a very particular context of the situation to reflect his/her intentions, purposes, world views and political positions. The sign makers, who are obviously positioned in favor of the Palestinian cause, have exploited the criterial characteristics manifested in Jewish settlement and the genocidal acts against the people of Gaza to transform the name 'Israel' into a verb in the past participle form as an apt signifier for the signified 'be cheated'. In the selection of the apt signifier, a metaphorical relation is selected by the sign-maker, who believes that the metaphor has the appropriate characteristics for the signified (Kress 1993). Taub (2004: 9) illustrates this criterial aptitude in connection with the coinage of new signs deriving from existing signs. She argues that the use of the new signs reflects the users' beliefs about the original sign. She exemplifies that 'dogged persistence' reflects our attitude

towards 'dogs' as patent and persistent.

The concept of motivated sign, which underscores the interplay between the social anchoring of signs, the individual contribution in their production and the signs' material appropriateness explained above, can account for sign mutability. According to Kress and Van Leeuwen (2001), changes in the linguistic system like the mutability of signs are motivated by the larger scale social, economic and political practices and are situated within the domain of ideology. Sign mutability explains the ideological positioning of the sign-maker. The ideological embedding of sign mutability can be a concretization of DeSaussure's notion of social force. This force is, however, not necessarily extended over a long period of time. Kress and Van Leeuwen (2001: 36) correlate the force of change in society with the intensity of change in modes of representation saying that "the intensity of the dynamic for change in practices increases the intensity of the dynamic of change in the various articulatory modes, and in the discourses which they articulate". Seen from this perspective, we can say that changes of attitudes and growing empathy towards the Palestinian cause have induced this coinage.

This section draws upon the Social Semiotic principle of motivated sign according to which signs are motivated by social forces and the sign-maker's interest. The latter chooses the materially and socially appropriate signifiers to express the intended meaning (Van Leeuwen 2005). In this process, ideology stands as a mediating element that can explain the choice of signs and their mutability. From this perspective, there is no neutral arbitrary sign and all signs are subject to criticism. Therefore, CDA rests on the principle of motivated sign to undertake a critical reading of signs. (Kress 1993). The following section will highlight how the iconic nature of GOT ISRAELED facilitates its ideological work adopting Peirce's principle of 'hypo-iconicity in iconic signs (Farias and Queiroz 2006).

5. Iconicity: Resemblance and metaphoric distortion

In the essay entitled 'on a new list of categories' in *Syllabus*, Peirce divides signs into *icons*, *symbols* and *indexes* (Peirce 1931). Indexes relate signs with the ideas they represent with a factual relation, for instance smoke as representation of fire. Symbols connect signs with the idea they represent by inferred characteristics as in white color symbolizing life. Icons relate signs with the objects they represent by shared qualities such as, phonological resemblance in *onomatopoeia*. Unlike DeSaussure, Peirce asserts that icons are not marginal signs but central to communication (Chu 2020). Peirce highlights the importance of icons saying "every assertion must contain an icon or a set of icons, or else must contain signs whose meaning is only explicable by icons (Peirce 1931 CP 1.158).

Peirce introduces hypo-iconicity as a conceptual framework to describe the essential features of iconic signs. These features are direct resemblance, schematization and metaphoric distortion. In resemblance, the sign has a direct straightforward resemblance with an object in real life (Chu 2018, 2020). Meta-

phoric distortion is a distancing in the expression of meaning between the sign and the object. The sign is placed in a new environment, which invites the reader to exert an interpretation of the sign but this interpretation does not prevent the reader from establishing and reconstructing the historical contextual connection with the object (Chu 2018, 2020). Schematization refers to the transformation of the icon into a pictural diagrammatic form to gain some degree of abstraction. Since the icon in this study is verbal, this feature of schematization is not applicable. This aligns with Chu's (2020) statement that the three features are not necessarily present together in iconic signs. Hypo-iconicity is a scale that describes how iconic signs operate within a spectrum from direct resemblance with the object/idea they represent to diagrammatic abstraction and creative/metaphorical distortion (Chu 2018, 2020).

In the present sign, the features of direct resemblance and metaphorical distortion are applicable and enhance its critical potential by inviting the reader to draw connections. At the resemblance level, by the combined influence of the phonological resemblance with the name, 'Israel' and the resemblance in the shared qualities of cheating according to the sign maker, the reader can make direct connections between the signifier GOT ISRAELED, the signified (taking something from the original owner) and the object/idea in real life 'Israel'. This triadic connection is accentuated by the fact that the sign is a closed one. This is illustrated in the paradigmatic connections this sign can/cannot establish with other signs. As example 1 shows, this sign imposes selection restriction constraints making paradigmatic substitutions of ISRAELED with other lexical items unacceptable as in b and c.

- (1) a- I got israeled.
b- *I got loved.
c- *I got stolen.
d- I got cheated/robbed.

These selection restrictions narrow down the signified to a sudden, negative and painful offense executed by an israeling force upon the 'subject', who is normally a human being taking a *goal* role. Due to this closedness, signifier and signified become mutually evocative. By virtue of the phonological resemblance between the sign and the name Israel, the reader can establish direct straightforward cognitive connections between the sign and Israel. Taub (2004: 21) stresses the link between resemblance and the reader's cognitive processing and critical reading saying that "there is no such thing as 'resemblance' or 'similarity' in the absence of an observer who makes a comparison: resemblance is not an objective fact about two entities but is a product of our cognitive processing" (Taub 2004: 21). These cognitive connections are likely to stereotype Israel as treacherous, which is obvious even among advocates of Israel. Steinberg (in Lebovic 2024) says in *Times of Israel* "the usage suggests that Israel inherently represents unjust appropriation and that offensive behavior typifies Israel. It is an antisemitic use of our people's ancient name".

At the level of metaphoric distortion, the idea of GOT ISRAELED is put in a casual, apolitical environment through the use of several techniques in the definition of the entry and in the example section. In the definition, apolitical participants are used, such as 'someone', 'they', and 'you'. In the example section, a casual usage example is given, in which someone is sitting in a table in a restaurant or a coffee shop and allows a person to share the table with him/her but at the end, this person asks him/her to leave because s/he has a meeting. The use of the table metaphor gives a pseudo-objective aspect to the lexicographic entry. The reader, however, can easily connect the sign with the political meaning and establish the inevitable connection with the historical context of the colonization of Palestine due to the resemblance feature of the sign explained above.

In this section, through applying the conceptual framework of hypo-iconicity, the analysis has shown that the sign displays two essential features of iconic sign, which are resemblance and metaphoric distortion. These features boost the critical potential of the sign and allow the reader to build direct straightforward connections between the sign and the political reality of the colonization of Palestine by Israel. In this study, iconicity is used as a distinct term from iconisation. The latter refers to the strategy of pooling attitudes around iconic signs, which is also a strategy of meaning-making aiming to expose the writers' political views and align the reader around them. This will be explained in details in the following section.

6. Iconisation: Ideation and bonding icon

Iconisation is a semiotic tool and an evolving research direction in SFL that examines the role of signs in "solidarity-building and affiliation" (Stenglin 2012: 125). It refers to the process through which a sign (symbol, icon or index) becomes a bonding icon, "a social emblem, which communities rally around or reject" (Stenglin 2012: 125). A bonding icon realizes the bonding through merging ideational and interpersonal meanings (Stenglin 2004). The ideational meanings connect to the *field* components of register in SFL (cf. Halliday and Matthiessen 2014), such as what the sign is both concretely and symbolically, the social activities associated with it and the underlying social processes (Stenglin 2008, 2012). Such processes can explain how signs are iconized. At the interpersonal level, the bonding is materialized through infusing attitudes to ideational meanings, which results in aligning people into communities and social networks with shared feelings and attitudes (Knight 2010). Turning a sign into a bonding icon is accompanied by a strong interpersonal charge with the amassing of interpersonal attitudes. Martin (2010) claims that ideational meanings are discharged while Stenglin (2012) contends that the ideational is weakened while the interpersonal is intensified and crystallized. The intensification of the interpersonal is mediated through several semiotic resources in verbal and multimodal forms (Wignell, Tan and O'Halloran 2016). Stenglin (2008) explores the expansion of the interpersonal in iconisation through the Appraisal System to gauge the expansion of attitudes

(affect, appreciation and judgement) as developed by Martin and White (2005).

To illustrate the iconisation process, flags, for instance, are bonding icons merging ideational and interpersonal meanings. Ideationally, the sign is associated with concrete component parts evoking its historical and cultural anchoring and is associated with different social activities such as hoisting flags during official ceremonies. Interpersonally, values get charged into this sign such as, pride, belonging, patriotism as well as homesickness, affiliating people into communities of shared values called citizens or countrymen.

The sign under examination is a typical bonding icon that merges ideational and interpersonal meanings. Ideationally, despite the surface apolitical definition discussed in the previous section, the sign recontextualizes the colonization of Palestine, the ethical illegitimacy in this colonization, Jewish settlements as well as the issue of violence associated with them. This ideational recontextualization has induced an interpersonal charging and an intensification of attitudes. The large community of internet users are rallied around a set of values such as the condemnation of the offender and sympathy with the victim. This amassing of attitudes has three manifestations:

- The wide resemiotisation of the sign inside the dictionary and in other mediated forms, playing two roles: being simultaneously a manifestation of iconisation and a semiotic tool fostering this process.
- The large-scale reactions to the sign through social media sharing and the large upvoting in *Urban Dictionary*. Zappavigna (2019) refers to this as an ambient affiliation through which the sign bonds internet users virtually.
- The use of semiotic resources in the definition to underscore the writers' attitude and the attempts to align the reader with it.

The current literature contends that a bonding icon triggers alternative reactions: communities either endorse an icon or reject it. In both situations, the icons are "supercharged and radiate specific values for communities to rally around" (Stenglin 2022: 6). *Bond-icons* evoke the communities' positive attitudes towards an icon, while *anti-cons* evoke rejection (Carr 2023, Stenglin 2022, Knight 2010). In both cases, bonding icons transcend differences and lead to unity and conformity. Stenglin (2012) argues that the Olympic torch is a bonding icon uniting people around the values of excellence, friendship and respect associated with the Olympic values to the point that the Nazi origin of the torch is totally eradicated. To my knowledge, there is a scarcity of research on a bonding icon triggering divergent attitudes across communities. The present paper contends that communities are rallied around the bonding icon according to two opposing directions. Pro-Palestine voices endorse the sign and align around the values it promotes. Pro-Israeli voices reject the sign through downvoting it in *Urban Dictionary*, and organizing social media campaigns for this purpose. The divergence of attitudes towards a bonding icon is expressed thanks to the affordances of online dictionaries and the crowdsourced ones, in particular.

To uncover the semiotic resources used by meaning-makers for the iconi-

sation of the sign, the analysis draws upon a toolkit from SFL to map the construction and aligning of attitudes. According to this theory, every situation of communication is an act of meaning (Halliday and Matthiessen 2014), which fulfills simultaneously three metafunctions: an ideational, an interpersonal and a textual. In the context of iconisation where ideational and interpersonal meanings are compressed, the two first functions are particularly relevant. They are defined as follows:

- The ideational function refers to the construal of experience through a configuration of participants, process types and optional circumstances. The participants are the represented participants who perform different participant roles depending on the process type. For instance, in the sentence, 'John hits Mary', 'John' and 'Mary' are the represented participants assuming the roles of *actor* and *goal* respectively. This meaning is mediated through the *material process*, 'to hit'. In counterpart, in 'John loves Mary', 'John' and 'Mary' are the represented participants assuming the roles of 'senser' and 'phenomenon' respectively. This meaning is mediated through the *mental process*, 'to love'.
- The interpersonal function is explored from the principle of interactive participants, which refers to the relation between interactants, who are the participants in a conversation or the writer, on the one hand and the reader, on the other (Kress and Van Leeuwen 1996/2021). Different resources mediate this relation in SFL such as, the articulation of positive and negative attitudes, which can be graduated by *intensification* or *weakening* (Martin and White 2005).

Table 1 introduces the resources used to map the iconisation of the sign under examination. It proposes a two-levelled analysis to gauge the coupling of ideational and interpersonal meanings.

Table 1: Ideational and interpersonal systems for the analysis of iconisation

	Ideational	Interpersonal
Participants and participant roles	Represented participants Process types and participant roles (Halliday and Matthiessen 2014)	Interactive participants (Kress and Van Leeuwen 1996/2021)
Ideational expansion and interpersonal charging	Metaphoric distortion (Peirce 1931)	Graduation (Martin and White 2005)

6.1 Represented participants and interactive participants

At a first level, ideationally, focus is laid on the identification of the represented participants, the participants introduced in the definition of the entry and in the

example. Emphasis is also put on the roles of represented participants (actor, goal, sayers, sensor) and mediating process types (*material*, *mental*, and *verbal*). Interpersonally, emphasis is put on the interactive participants, the writers of the entry, and the readers and how they connect to each other.

6.1.1 Represented participants, their roles and process types

The represented participants can be divided into two broad categories: the israeling participant (oppressor) and the israeled (victim). The oppressor is associated with the third person in the singular and plural, 'someone' is used in 20 instances and 'they' in 17. The victim is frequently associated with the first person singular and the second person plural; 'I' is employed 19 times and 'you' 17. This distribution is meant to identify with the victim, who is either the speaker or the reader. On the contrary, the israeling participant is distanced through the use of the third person to reflect the authors' negative attitude to them. The shift from the third person singular 'he/someone' to the plural 'they' indicates that the israeling act has evolved from the act of a single person to the work of a group of people. In addition to the use of personal pronouns, the two categories of participants are referred to with names. In seven entries, names with apolitical orientation (Elizabeth, Peter, John, Robert and Nick) are used to give a casual dimension to the entries. As the resemiotisations progress in time, proper names with obvious political, cultural and symbolic connotations are used. Names such as, Benjamin, Jacob, Hananya versus Ishmail, Ali, Maryem and Hamad obviously recontextualize the conflict between Jews and Arabs. In this conflict, the writers of the entries are sided by the Palestinians, who are described as the 'rightful owner'. On the contrary, nouns and names such as, 'trespasser' and 'Barbara' (a name in Greek mythology to refer to a foreigner or a stranger) are used to refer to the israeling participant.

Table 2 shows that the israeling participant has functioned more frequently as the subject of the sentence assuming active participant roles mediated by verbal, material and mental process types (143 versus 19). The material processes with an actor participant role are the most frequent (70 material processes). They are realized by the verbs 'steal' and 'take' (29 processes) aiming to construct the idea of unlawful taking. This culminates in the use of 'to israel' in the active voice, which is exclusively used in the resemiotisations (12 processes). The clear straightforward association of this with the war on Palestine is made through the use of the verb 'to kill' (3 times) as in "when you resist they kill you for it and take everything else you own" (ISRAELED, 5 November 2023). The Ideological anchoring of the entry is also evident through associating GOT ISRAELED with the victimhood narrative in Jewish communities (Markiewicz and Sharvit 2021). This is mentioned in five entries as in "to act like the victim 24/7, but is able to steal things that belong to someone else" (ISRAELED, 5 November 2023).

The verbal processes come second in frequency with 63 occurrences. The verbs 'ask' and 'claim' are the most frequent (14 and 15 occurrences respectively).

These processes are chronologically sequenced to mean that the israeling act started with asking someone to share something with him/her and then claiming it as one's property as in "when someone asks to share something of yours and then claims it's become theirs and takes it for themselves" (GOT ISRAELED, 28 October 2023).

As opposed to the frequent occurrences of the israeling participant in a subject position with actor and sayer participant roles, the israeled participant assumes these roles only 19 times, which are often associated with the verbs 'share', 'allow' and 'give' (six verbs). The political nature of the entries is also highlighted through the use of the verbs of 'refuse' and 'resist' as in "when you resist, they kill you for it and take everything else you own" (ISRAELED, 5 November 2023). As Table 2 shows, the israeled participant is more frequently associated with the goal participant role in both subject and object positions. In all the occurrences of subject as goal, the israeled participant is used with the verb 'to israel' in the passive voice or with the mutative 'got'.

Table 2: Participant roles in the entries

Participant roles	Israeling participant	Israeled participant
Subject as actor/sayer/ carrier	143	19
Subject as goal	0	30
Object as goal	2	30

6.1.2 Interactive participants

Writers of the entry are portrayed as articulate and biased participants in the different entries by means of several communication strategies. In 19 instances, the writer is identified as the victim as in "I was sitting at my table I think I just GOT ISRAELED". The use of pseudonyms such as 'Israeled', 'an Israeled man' indicate the direct involvement of the writer. Another manifestation of the writers' bias is the use of pseudonyms with clear political affiliations such as 'October Revolution' 'FreeFelestin'.

Writers establish direct bonds with the reader. This is illustrated in 17 entries, where he/she addresses the reader directly and identifies him/her as the victim as in "when you allow someone to use a thing you own or sit in a place reserved for you and then they claim it's theirs all along" (ISRAELED, 21 October 2023). The reader can express his/her attitude towards these entries thanks to upvoting and downvoting. These reactions reflect that the sign triggered divergent attitudes among internet users (almost 600000 reactions from 21 October 2023 to 11 October 2024). It is noticeable that the total number of upvoting exceeds downvoting (387405 versus 201628).

To sum up, the ideational choices (participant roles and process types) indicate that the israeling participant is introduced as the major participant in the

construal of experience, who is taking what originally belongs to the israeled participant. However, through the use of personal pronouns, the writers identify with the israeled participant and distance themselves from the israeling one. This strategy leads to the interpersonal charging of the sign with positive attitudes towards the victim and negative ones towards the oppressor. Interpersonally, this charging is translated into positioning the victim as teller of the story. This highlights the causal connection between the two metafunctions; the reality of colonization has amassed positive interpersonal attitudes such as sympathy and compassion towards the victim. This is congruent with the analysis of interactive participants, which also highlights the interpersonal charge of the sign manifested in the reactions of the readers.

6.2 Metaphoric distortion and graduation

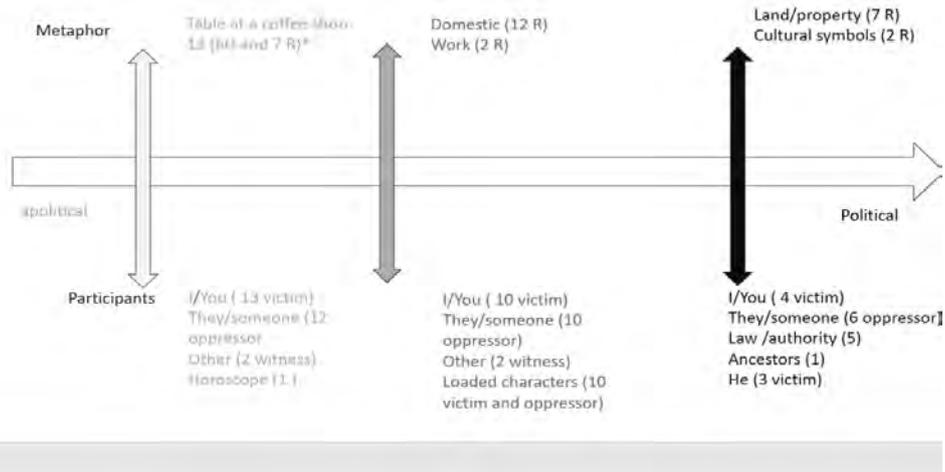
At a second level, the analysis of metaphoric distortion reveals the ideational expansion of the sign through the escalating recontextualization of reality as the resemiotisations unfold. The graduation analysis underscores the interpersonal charging of the bonding icon.

6.2.1 Metaphoric distortion

Metaphoric distortion is a central feature of iconic signs. The analysis focusing on the original entries (those written on 21 October 2023) done in section 5 has shown that the 'table metaphor', a casual example used as a metaphoric explanation of GOT ISRAELED, does not prevent the reader from establishing straightforward connections with the colonization of Palestine. This section concentrates on the metaphoric distortions in the resemiotisations to underscore the progressive recontextualizations of this colonization as the resemiotisations progress in time. The analysis of metaphoric distortion covers shifts at three areas:

- Nature of metaphor across resemiotisations
- Participants' identity
- Use of process types

As indicated in Figure 3, the use of metaphors shifts towards a more political orientation as the resemiotisations unfold. The original entries use a casual metaphor of a table in a restaurant or a coffee shop. It is accompanied by a humorous/sarcastic tone as in "is this an episode of Seinfeld or am I getting israeled?" (ISRAELED, 21 October 2023) or "my horoscope said that I should be sitting at this table" (ISRAELED, 22 October 2023). As resemiotisations progress, metaphors relating to the domestic and work domains are used such as israeling one's t-shirt, pen, ice-cream, bicycle or project. The political orientation becomes more evident in the last resemiotisations. The metaphors used become those of israeling land/property and cultural symbols such as Falafel and tangerines.



*O= original entry R= resemiotisation

Figure 3: Metaphoric distortion across entries

This incremental political shift is also mirrored in the use of participants. In the original entries, the participants are consistently used with 'I' and 'you' representing the israeled and 'someone' and 'they' representing the israeling participants. In the resemiotisations with the domestic/work metaphors, in addition to the participants mentioned above, others like mother, sister and colleagues are used. Most importantly, pairs of participants with cultural connotations are used, such as Ali versus Jacob and Ishmail versus Barbara. These pairs not only reflect the cultural orientation of the entry but also point out the conflict between the two cultures. In the property-related metaphors, the use of participants becomes less consistent ('he' is used for israeling and israeled participants). More importantly, intervening participants denoting power and authority are used to show that the israeling act is helped by power as in "when a person tells you that your property is theirs ... and demands you just give this property to them, and if you refuse, they take it by force and the law will somehow be on their (israeling) side" (ISRAELED, 22 October 2023). The political dimension is, in particular, emphasized in one resemiotisation, which depicts a conversation between a settler and the original owner of the land, in which the former claims the land belonged to his ancestors "my ancestors were here 3000 years ago. If I do not take your house, someone else will" (ISRAELED, 28 October 2023). This statement is actually identical to a previous utterance in a real-life video reporting a conversation between a Jewish settler and a Palestinian woman in Cheikh Jarrah area in 2021 (AJ+ 2021).

The evolving political orientation is also evident in the use of process types.

Though verbs like 'claim' 'take' 'steal' and 'kill' are employed in the original metaphor, their number increases in the domestic and property metaphors (for instance 'claim' evolves from four to seven and 'steal' from one to seven). The verb 'act like a victim' was scarcely used with the table metaphor as opposed to six uses in the other metaphors.

6.2.2 Graduation

Graduation is a system within the Appraisal Theory of Martin and White (2005: 135-155), which maps evaluation in language. It covers, alongside with *graduation*, *attitude* and *engagement*. Graduation is anchored within SFL's interpersonal metafunction and includes the resources for grading evaluation through *up-scaling* and *down-scaling*. As shown in Figure 4, graduation functions along two axes of evaluation, which are *force* and *focus*. Scalability along the axis of force covers grading according to *intensification* and *quantification*. Graduation along the axis of focus refers to grading according to *prototypicality*. The first sub-system of graduation includes evaluation relative to the degree of intensity as well as quantity. Assessment of degree of intensification can operate at the level of quality as in 'very kind' and 'quite helpful' and at the level of 'processes' 'this greatly surprised me' and 'you slightly disturbed me'. Graduation in quantification covers amount (size and number) and extent (time and space) as in 'a short break' and 'a large field'. Graduation according to prototypicality (focus) delineates the degree the phenomenon matches the prototypical features of a semantic category as in 'a true lover' (sharpening) and 'kind of lover' (softening).

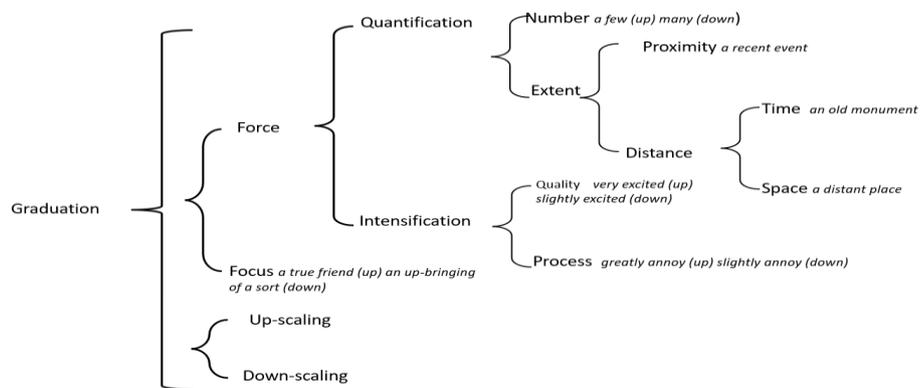
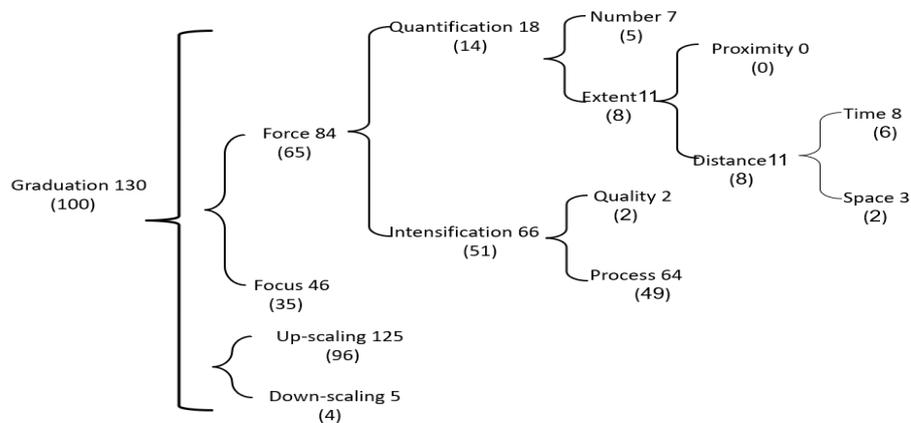


Figure 4: System network for graduation adapted from Martin and White (2005: 154)

Since graduation is concerned with grading experiences and attitude, this framework can be useful in understanding the iconisation of signs. Graduation can

reveal the resources the authors used in charging the sign with values in order to orient the reader. The entries include 130 graduation instances at an average of four graduations per entry. As shown in Figure 5, the direction of grading is predominantly up-scaling. Force-based graduation is more frequent than focus-based one (84 versus 46).

Graduation in terms of force using the sub-systems of intensity and quantification interacts with the up-scaling of attitude, which "frequently acts to construe the speaker/writer as maximally committed to the value position being advanced and hence as strongly aligning the reader into that value position" (Martin and White 2005: 152). Intensification of qualities by grading adjectives and adverbs are typically attitudinal. They are, however, rarely used in the entries under examination, which may be explained by the fact that clearly inscribed attitudes are incongruent with the genre of dictionary entries. On the contrary, intensification by processes is very frequent with 66 instances. This intensification is not typically attitudinal: up-scaled verbs are used to convey the invoked attitude of the authors towards the action of both israeling participants as in 'kick out the rightful owner' "get you out by force" (ISRAELED, 21 October 2023). Grading of processes is also used to convey the degree of powerlessness of the israelied as in "you can't resist it whatsoever" (To israel, 25 October 2023).



() The figures between parentheses refer to the percentage of each graduation resource.

Figure 5: Frequency of graduation resources in the entries

Quantification resources are used to convey the extent of the israeling act in terms of taking the victim's property as in "and take **everything else** you own" (ISRAELED, 5 November 2023). Quantification covering extent in time is used to highlight the religious anchoring of the israeling act "when someone asks for something from you and then **overtime** starts claiming that thing was **always**

theirs" (ISRAELED, 5 November 2023 emphasis added). The authors draw connection with the Jewish propaganda of promised land as "the act of taking something from someone else for some reason thousands of years ago" (ISRAELED, 28 October 2023). The use of these quantification resources is concentrated in the entries recontextualizing the religious conflict.

Graduation under focus sub-system amounts to 46 instances representing 35% of all graduation resources. They are attitudinal terms construing the speakers' attitude towards a situation or a fact, which are explicitly attitudinal as in "it can mean causing **harm** and **destruction** in a **tyrannical way**" (TO ISRAEL, 25 October 2023 emphasis added) or in "a **sickly exaggerated desire** to have something that belongs to someone else" (ISRAELING, 24 October 2023 emphasis added). These graduation resources "indicate maximal investment by the authorial voice in the value position (either negative or positive) being advanced and hence to strongly align the reader into the value position being advanced." (Martin and White 2005: 139)

To sum up, at the ideational level, metaphoric distortion is used to illustrate the shift into more politicized metaphors. This shift underscores the authorial voice, which is realized with the use of graduation resources meant to accentuate the interpersonal charge of the bonding icon. This finding aligns with Stenglin (2008) who emphasizes the coupling of ideational and interpersonal meanings to charge values into the Olympic flame. She demonstrates this through combining the ideational activities of lighting the Olympic flame with the attitudes of affect, appreciation and judgement evoked by these activities. The present study uses the coupling of metaphoric distortion and graduation because these systems are more congruent with the genre of dictionary entries where invoked attitudinal choices are genre-appropriate.

7. Conclusion

This paper has proposed to link Lexicography and CDA by demonstrating that the semantic and lexical descriptions of new dictionary entries should be supplemented with a critical reading. As shown by the study, the ideology-infused entries constitute interpersonally-charged semiotic acts introducing the sign-maker's world view and inducing reactions from the reader. This was illustrated by GOT ISRAELED, a new entry in the crowdsourced *Urban Dictionary* that was heavily resemiotised inside the dictionary reaching 36 entries.

The sign was analyzed at three levels: coinage, iconicity and iconisation drawing upon Semiotic and Social Semiotic frameworks. The first level of analysis shows that the sign is a motivated sign showing the correspondence between structural features, social anchoring and sign makers' intentions. At the second level, Peirce's (1931) principles of hypo-iconicity underscore the critical potential of the sign through drawing straightforward links with the colonization of Palestine. Third, this paper has exploited the Social Semiotic principle of iconisa-

tion to describe the transformation of the sign into a bonding icon rallying people around the writers' world view (Stenglin 2012).

This study has contributed to the current literature on the relatively new area of critical lexicography by introducing a framework for the analysis of signs with a semiotic action. This framework illustrated in Figure 6 represents a two-levelled analysis drawing upon tools from Critical Lexicography (Chen 2019), Semiotics (DeSaussure 1916, Peirce 1931) and Social Semiotics (Halliday and Matthiessen 2014, Martin and White 2005). The first level is a contextual analysis emphasizing the generic structure of the entry, the context of genre production and explores the specific political context of the entry coinage.

The second level of analysis comprises three steps showing that the potential of signs to exert action can be viewed on a spectrum ranging from the coinage of motivated signs to the design of signs building real and distorted connections with reality (iconicity). At the highest level of this spectrum is iconisation, in which the sign maker exploits ideational and interpersonal resources to turn the sign into a bonding icon. This level is reached when the sign is largely resemiotised in different mediated forms. The aims of resemiotisations are to construct a set of values around the iconized sign.

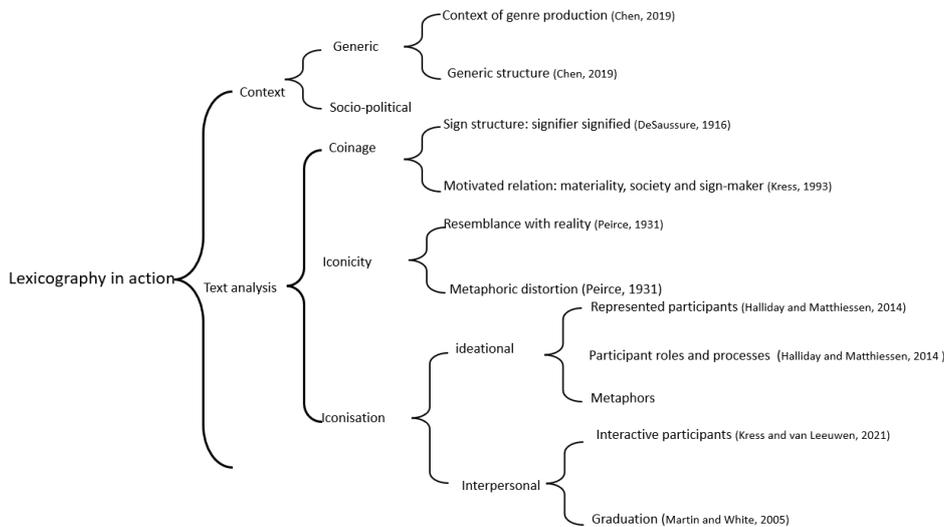


Figure 6: A Framework for the analysis of lexicographic entries in action

This framework has been used to analyze dictionary entries but it is applicable to other signs in language and other semiotic systems. Multimodality can promote semiosis in action at the levels of coinage and iconicity as well as iconisation because it can help the expansions of the sign's meaning. This can be a promising topic for future research.

References

- AJ+. 2021. If I Don't Steal [Your Home] Someone Else Will. Israeli Settler Justifies Forcible Takeover. [video]. YouTube, 5 May 2021.
<https://www.youtube.com/watch?v=KNqozQ8uaV8>.
- Bergenholtz, H. and R.H. Gouws. 2012. What is Lexicography? *Lexikos* 22: 31-42.
- Bertulesi, C. 2023. Chinese Lexicography and the Critical Analysis of Lexicographical Discourse. Zucheri, S. (Ed.). 2023. *Studies on Chinese Language and Linguistics in Italy*: 95-116. Bologna: Bologna University Press.
- Björkqvall, A. and A.-M. Karlsson. 2011. The Materiality of Discourses and the Semiotics of Materials: A Social Perspective on the Meaning Potentials of Written Texts and Furniture. *Semiotica* 187(1-4): 141-165.
- Carr, G. 2023. *With Respect to Consent: The Language of Sex Education*. Unpublished Ph.D. Dissertation. Sydney: University of Sydney.
- Chen, W.G. 2015. Bilingual Lexicography and Recontextualization: A Case Study of Illustrative Examples in a New English-Chinese Dictionary. *Australian Journal of Linguistics* 35(4): 311-333.
- Chen, W.G. 2019. Towards a Discourse Approach to Critical Lexicography. *International Journal of Lexicography* 32(3): 362-388.
- Chu, Y. 2018. Hypoiconicity in the Architecture of Suzhou: Authentic Resemblance, Diagrammatic Reduction, and Metaphoric Displacement. *Social Semiotics* 30(1): 114-132.
[doi:10.1080/10350330.2018.1531548](https://doi.org/10.1080/10350330.2018.1531548).
- Chu, Y. 2020. On the Hypoiconic Structure of Cartoons. *Social Semiotics* 32(2): 262-278.
<https://doi.org/10.1080/10350330.2020.1756587>.
- Coleman, J. 2004. *A History of Cant and Slang Dictionaries*. Oxford: Oxford University Press.
- DeSaussure, F. 1916. *Course in General Linguistics*. (Edited by Bally, C. and A. Sechehaye in collaboration with A. Riedlinger. 1966. Translated, with an Introduction and Notes by Wade Baskin.) New York/Toronto/London: McGraw-Hill.
- Fairclough, N. 1995. *Critical Discourse Analysis: The Critical Study of Language*. London: Longman.
- Farias, P. and J. Queiroz. 2006. Images, Diagrams, and Metaphors: Hypoicons in the Context of Peirce's Sixty-six Fold Classification of Signs. *Semiotica* 162(1-4): 287-308.
- Fuertes-Olivera, P.A. and S. Tarp. 2022. Critical Lexicography at Work: Reflections and Proposals for Eliminating Gender Bias in General Dictionaries of Spanish. *Lexikos* 32(2): 105-132.
- Gouws, R.H. 2014. Article Structures: Moving from Printed to e-Dictionaries. *Lexikos* 24: 155-177.
- Halliday, M.A.K. and C.M.I.M. Matthiessen. 2014. *Halliday's Introduction to Functional Grammar*. Fourth edition. Oxon: Routledge.
- Hodge, R. and G. Kress. 1988. *Social Semiotics*. Cambridge: Polity.
- Holdcroft, D. 1991. *Saussure: Sign, System and Arbitrariness*. Modern European Philosophy. Cambridge: Cambridge University Press.
- Hornscheidt, A. 2008. A Concrete Research Agenda for Critical Lexicographic Research within Critical Discourse Studies: An Investigation into Racism/Colonialism in Monolingual Danish, German, and Swedish Dictionaries. *Critical Discourse Studies* 5(2): 107-132.
- Kachru, B.B. 1995. Afterword: Directions and Challenges. Kachru, B.B. and H. Kahane (Eds.). 1995. *Cultures, Ideologies, and the Dictionary: Studies in Honor of Ladislav Zgusta*: 417-424. Tübingen: Max Niemeyer.

- Kirkness, A.** 2004. Lexicography. Davies, Alan and Catherine Elder. 2004. *The Handbook of Applied Linguistics*: 54-81. Oxford: Blackwell.
- Knight, N.K.** 2010. Wrinkling Complexity: Concepts of Identity and Affiliation in Humour. Bednarek, M. and J.R. Martin (Eds.). 2010. *New Discourse on Language: Functional Perspectives on Multimodality, Identity, and Affiliation*: 35-58. London: Continuum.
- Kress, G.** 1993. Against Arbitrariness: The Social Production of the Sign as a Foundational Issue in Critical Discourse Analysis. *Discourse and Society* 4(2): 169-193.
- Kress, G.** 2010. *Multimodality: A Social Semiotic Approach to Contemporary Communication*, London: Routledge.
- Kress, G. and T. van Leeuwen.** 1996/2021. *Reading Images: The Grammar of Visual Design*. London: Routledge.
- Kress, G. and T. van Leeuwen.** 2001. *Multimodal Discourse: The Modes and Media of Contemporary Communication*. New York: Hodder Arnold.
- Lebovic, M.** 2024. Is *Urban Dictionary*'s New Entry 'to Get Israeled' an Antisemitic Trope? *The Times of Israel*, 11 January 2024.
<https://www.timesofisrael.com/is-urban-dictionary-s-new-entry-to-get-israeled-an-antisemitic-trope-adl-weighs-in>
- Ledin, P. and D. Machin.** 2018. Multi-modal Critical Discourse Analysis. Flowerdew, J. and J.E. Richardson (Eds.). 2018. *The Routledge Handbook of Critical Discourse Studies*: 60-76. London/New York: Routledge.
- Linell, P.** 1998. Discourse across Boundaries: On Recontextualizations and the Blending of Voices in Professional Discourse. *Text* 18(2): 143-157.
- Liu, X.** 2015. Multimodal Definition: The Multiplication of Meaning in Electronic Dictionaries. *Lexikos* 25: 210-232.
- Markiewicz, T. and K. Sharvit.** 2021. When Victimhood Goes to War: Israel and Victim Claims. *Political Psychology* 42(1): 111-126.
- Martin, J.R.** 2010. Semantic Variation: Modelling Realisation, Instantiation and Individuation in Social Semiosis. Bednarek, M. and J.R. Martin (Eds.). 2010. *New Discourse on Language: Functional Perspectives on Multimodality, Identity, and Affiliation*: 1-34. London: Continuum.
- Martin, J.R. and P.R.R. White.** 2005. *The Language of Evaluation. Appraisal in English*. New York: Palgrave Macmillan.
- Moalla, D.** 2023. The Resemiotisation of Chokri Belaid's Talk before his Assassination: Roles of Sign-maker Interest and Re-reading Forces. *Social Semiotics* 33(4): 808-840.
DOI: 10.1080/10350330.2021.1949245.
- Peirce, C.S.** 1907. Pragmatism. Houser, N. and C.J.W. Kloesel (Eds.). 1998. *The Essential Peirce. Vol. 2: Selected Philosophical Writings (1893-1913)*: 398-433. Bloomington: Indiana University Press.
- Peirce, C.S.** 1931. *The Collected Papers of Charles Sanders Peirce*. Hartshorne, C., P. Weiss and A.W. Burks (Eds.). 1966. Cambridge: Harvard University Press. [Reference to Peirce's papers will be designated CP followed by volume and paragraph number.]
- Pettini, S.** 2023. The Language of Ethnic Conflict in English Online Lexicography: Ethnophaulisms in "powered by Oxford" Lexico.com. *Lexikos* 33: 299-317.
- Ro, C.** 2019. How Linguists Are Using *Urban Dictionary*. *Daily JSTOR*, 13 November 2019.
<https://daily.jstor.org/how-linguists-are-using-urban-dictionary/>
- Stenglin, M.** 2004. *Packaging Curiosities: Towards a Grammar of Three-dimensional Space*. Unpublished Ph.D. Dissertation. Sydney: University of Sydney.

- Stenglin, M.** 2008. Interpersonal Meaning in 3D Space: How a Bonding Icon Gets its 'Charge'? Unsworth, L. (Ed.). 2008. *Multimodal Semiotics: Functional Analysis in Contexts of Education*: 50-66. London: Continuum.
- Stenglin, M.** 2012. 'Glocalisation': Exploring the Dialectic between the Local and the Global. Bowcher, W.L. (Ed.). 2012. *Multimodal Texts from Around the World: Cultural and Linguistic Insights*: 123-145. Houndmills, Basingstoke, Hampshire/New York: Palgrave Macmillan.
- Stenglin, M.** 2022. Binding and Bonding: A Retrospective and Prospective Gaze. Caldwell, D., J.S. Knox and J.R. Martin (Eds.). 2022. *Applicable Linguistics and Social Semiotics: Developing Theory from Practice*: 309-324. London: Bloomsbury.
- Strand, T.** 2013. Peirce's Rhetorical Turn: Conceptualizing Education as Semiosis. *Educational Philosophy and Theory* 45(7): 789-803.
- Taub, S.F.** 2004. *Language from the Body. Iconicity and Metaphor in American Sign Language*. Cambridge: Cambridge University Press.
- Urban Dictionary.** 2023–2024. *Israeled*.
<https://www.urbandictionary.com/define.php?term=Israeled>
- Urban Dictionary.** 2024. *Homepage*.
<https://www.urbandictionary.com/define.php?term=homepage>
- Van Leeuwen, T.** 2005. *Introducing Social Semiotics*. London/New York: Routledge.
- Wignell, P., S. Tan and K.L. O'Halloran.** 2016. Violent Extremism and Iconisation: Commanding Good and Forbidding Evil? *Critical Discourse Studies* 14(1): 1-23.
- Zappavigna, M.** 2019. Ambient Affiliation and #Brexit: Negotiating Values about Experts through Censure and Ridicule. Koller, V., S. Kopf and M. Miglbauer (Eds.). 2019. *Discourses of Brexit*: 48-68. London: Routledge.

Reconceptualising Contemporary Lexicography: Emotional Intelligence and the Sustainable Lexicographic Modelling of the Modal Verb *moći*

Milena Mrdak-Mićović, *Department of English Language and Literature,
Faculty of Philology, University of Montenegro, Nikšić,
Montenegro (milenam@ucg.ac.me)*
(<https://orcid.org/0009-0008-7135-9228>)

and

Nataša Jovović, *Department of Montenegrin Language and South Slavic
Literature, Faculty of Philology, University of Montenegro,
Nikšić, Montenegro (natasaj@ucg.ac.me)*
(<https://orcid.org/0009-0003-5857-3936>)

Abstract: The article examines the modal verb *moći* ("can"/"to be able to") as described in dictionaries currently in active use within the Montenegrin linguistic context, with particular emphasis on its significance in both digital and interpersonal communication in contemporary language. Building upon the foundational ideas of the lexicographic project *Algemeen Nederlands Woordenboek* (ANW) (Fuentes-Olivera 2024, following Colman 2016), the authors propose that the analysis and description of lexical units — described on the example of this particular modal verb — should be preceded by a comprehensive analysis in the domain of the emotional impact on interlocutors during conversation. This approach would influence the development of dictionaries that go beyond simply defining lexical units in various contexts. Instead, they would offer guidance on using words in emotionally calibrated communication to effectively convey and elicit the intended emotional impact within a given context. Thus, the aim of this paper is to highlight the importance of emotional intelligence and regulation through the example of the modal verb *moći*, and to offer recommendations for more effective management of emotional tone across diverse communicative environments. In this context, contemporary dictionaries would play a pivotal role in enhancing the quality and effectiveness of communication, not only by offering precise semantic distinctions, but also by incorporating pragmatic and affective guidance, thereby ensuring that lexicography remains a sustainable and socially responsive discipline. This multidimensionality demands not only lexical precision but also contextual thoughtfulness and emotional attunement to achieve communicative effectiveness, making lexicography a sustainable discipline that remains at the heart of all linguistic research, where it rightfully belongs.

Keywords: LEXICOGRAPHY, MODAL VERBS, DIGITAL COMMUNICATION, DICTIONARIES, EMOTIONAL INTELLIGENCE, DISCOURSE ANALYSIS

Opsomming: Die herkonseptualisering van kontemporêre leksikografie: Emosionele intelligensie en die volhoubare leksikografiese modellering van die modale werkwoord *moći*.

Hierdie artikel ondersoek die modale werkwoord *moći* ("kan"/"om in staat te wees tot") soos dit beskryf word in woordeboeke wat tans aktief binne die Montenegrynse linguistiese konteks gebruik word, met spesifieke klem op die belang daarvan in beide die digitale en interpersoonlike kommunikasie in kontemporêre taalgebruik. Deur voort te bou op die fundamentele idees van die leksikografiese projek *Algemeen Nederlands Woordenboek* (ANW) (Fuertes-Olivera 2024, in navolging van Colman 2016), stel die outeurs voor dat die analise en beskrywing van leksikale eenhede — soos beskryf met hierdie spesifieke modale werkwoord as voorbeeld — voorafgegaan moet word deur 'n omvattende analise van die domein van die emosionele impak op gespreksgenote tydens 'n gesprek. Hierdie benadering sal die ontwikkeling van woordeboeke wat wyer strek as slegs die definiëring van leksikale eenhede in verskillende kontekste, beïnvloed. In plaas daarvan sal hulle leiding bied in die gebruik van woorde in emosioneel gekalibreerde kommunikasie om die beoogde emosionele impak effektief binne 'n gegewe konteks oor te dra en te ontlok. Die doel van hierdie artikel is dus om die belangrikheid van emosionele intelligensie en die regulering daarvan aan die hand van die modale werkwoord *moći* as voorbeeld te belig, en om aanbevelings vir die effektiewer bestuur van emosionele toon oor diverse kommunikatiewe omgewings te maak. Binne hierdie konteks sal kontemporêre woordeboeke 'n sentrale rol in die verbetering van die kwaliteit en effektiwiteit van kommunikasie speel, nie net deur eksakte semantiese onderskeidings aan te bied nie, maar ook deur die inkorporering van pragmatiese en affektiewe leiding, om sodoende te verseker dat die leksikografie 'n volhoubare en sosiaal reaktiewe dissipline bly. Hierdie multidimensionaliteit vereis nie net leksikale presisie nie, maar ook kontekstuele bedagsaamheid en emosionele aanvoeling om kommunikatiewe effektiwiteit te bereik, om die leksikografie 'n volhoubare dissipline te maak wat steeds die kern van alle linguistiese navorsing vorm, waar dit regmatig hoort.

Sleutelwoorde: LEKSIKOGRAFIE, MODALE WERKWOORDE, DIGITALE KOMMUNIKASIE, WOORDEBOEKE, EMOSIONELE INTELLIGENSIE, DISKOERSANALISE

1. Introduction

The objective of this article is to identify the ways in which the modal verb *moći* ("can"/"to be able to") is described in dictionaries currently in active use within the Montenegrin-speaking region, with a particular focus on the role this modal plays in both digital and interpersonal communication among speakers of contemporary language. Drawing on the core ideas of the *Algemeen Nederlands Woordenboek* (ANW) lexicographic project (Fuertes-Olivera 2024, following Colman 2016), the findings of this study suggest that the analysis and description of lexical units — specifically, the modal verb *moći* — should be preceded by research concerning their emotional impact on speakers. Such an approach would carry important implications for the development of dictionaries that not only describe the contemporary language but also offer practical recommendations for the use of a particular lexical unit in present-day communication. The case study carried out for the purpose of this article was perceived as an opportunity to illustrate how modern lexicography must move beyond static definitions and

instead grapple with meaning as a dynamic interplay of sense, affect, and inter-subjective relation within a given context. Therefore, the objective of the research was not solely to ascertain what this modal *means*, but to explore how it *is felt* — to trace the contours of its experiential resonance in lived language. Consequently, lexicography could evolve into a richer and more sustainable resource — one that not only explains word meanings but also supports users in understanding the emotional resonance of words in communicative contexts. This would enable more effective and responsible language use. The aim of this paper is to shed light on the role of the modal verb *moći* in emotional regulation and to provide recommendations for more effective management of emotional tone across various communicative environments, where contemporary dictionaries are expected to be of central importance in enhancing the quality and efficiency of communication.

Data obtained through both qualitative and quantitative analysis provided insights into the ways speakers of different age groups use the modal verb *moći* to shape emotional tones, with the aim of identifying its connotative meanings and pragmatic functions within various discourses. The results may significantly contribute to the development of contemporary language dictionaries and handbooks by offering more precise definitions and usage examples that illuminate how modal verbs and modalized expressions are used to either soften or intensify emotional responses in communication. This emphasis on documenting and further analyzing linguistic forms within a lexicographic framework will support the advancement of dictionary resources, laying the groundwork not only for the creation of more comprehensive dictionaries but also for handbooks that could play a notable role in media literacy. Such resources would be of particular value to speakers of all ages, especially in contemporary communication contexts — often written and emotionally charged — where individuals may feel vulnerable.

We align ourselves with linguists such as Šubarić and Đurčević (2023) and Nenezić (2024), who note that scholars like Apresjan (2010) and Topolinjska (2002: 33) emphasize the necessity of mutual coherence in terms of both content and data collection methods. This kind of integration presupposes close collaboration among authors, grounded in shared theoretical foundations and the consistent application of principles of "identification, classification, and interpretation of linguistic units" (ibid.). However, as Topolinjska (2002: 33) and Bartsch (1987) observe, such interdisciplinary coherence remains rare in linguistic practice. Therefore, the aim of this paper is to demonstrate that the expanded interpretation not only of modal verbs and modalized expressions, but of all lexical units, could significantly assist contemporary language users — regardless of whether the language is their mother tongue — in using it in a more deliberate and emotionally responsible manner.

2. The concept of contemporary lexicography and an integrated database

Since the 1990s, linguists have increasingly advocated for a model of lexicogra-

phy that extends beyond traditional descriptive language analysis, arguing that its development must adopt an interdisciplinary approach — incorporating insights from psychology, sociology, and cognitive science. Sinclair (1991) emphasized the need to link semantics with cognitive aspects rooted in cultural studies, while Anna Wierzbicka (1992) went a step further by exploring universal linguistic concepts, considering them essential for a deeper understanding of language. Her perspective appears to align with that of Geeraerts (2010), whose research on word meaning — with particular focus on cognitive and cultural contexts — presents an integrated approach to lexicology as not only logical but also inevitable and necessary.

In this context, a significant contribution comes from Fuertes-Olivera (2024) in his work *Making Lexicography Sustainable: Using ChatGPT and Reusing Data for Lexicographic Purposes*, where he argues that

we must go beyond the language-centered lexicographic tradition that dominates current thinking and focus instead on new thinking centered on increasing lexicographic productivity and using technologies that (a) adopt a broad concept of lexicographic data, (b) speed up the lexicographic process, (c) save time and reduce costs, (d) facilitate direct cognitive processing, e.g. by machines, and (e) allow the individualization of data as units of consumption and sale. In particular, we must critically examine the benefits and drawbacks of the different practices on offer. (Fuertes-Olivera 2024: 140)

The author grounds his innovative ideas in the work of Chomsky et al. (2023), Colman (2016), and De Schryver (2023), offering a fresh perspective on lexicography by approaching it as a modern scientific discipline. Although still evolving, lexicography is envisioned here as a field that must serve the needs of contemporary language users. It is upon these foundational claims that we base the present study.

One of the potential aims of modern lexicography could be the development of integrated and interactive databases that include lexical units enriched with information about their emotional and contextual appropriateness. Such databases would not merely identify the semantic meanings of lexical items, but would also enable users to understand not only the contextual framework traditionally provided in dictionary entries, but also the emotional tone associated with particular expressions. This emotional nuance — especially shaped by modalized expressions and lexical choices — can significantly influence how an interlocutor interprets a given message, often in subtle and highly context-dependent ways.

3. Understanding words and their impact in the communication process

Given that the Latin root of the word *communicare* means "to transmit" or "to share," it can be said that every spoken word has the potential not only to convey information but also to transmit emotion and provoke a broad spectrum of emo-

tional reactions in the interlocutor. These reactions largely depend on the context of use, tone of speech, and a broader framework of cultural, social, and psychological factors (Calvo and D'Mello 2010; Hall 1976; Tubbs 2012; Van Dijk 2008, 2009). Taking into account the findings of Anna Wierzbicka (1999), who argues that language is not merely a neutral carrier of meaning but reflects culture-specific patterns tied to speakers' emotional perceptions, it becomes evident that the complex dynamics of communication must inevitably influence the lexicographic practice of a language, which should follow the evolution of that language.

This argument is further supported by more recent research (Pavlenko 2014; Mesquita et al. 2021), proving that languages structure emotional expressions in different ways that may either facilitate or hinder intercultural understanding. Lexicography, therefore, deserves particular attention in the process of language description, as it is not only a discipline that documents linguistic resources but also one that shapes the full perception of word meanings and their usage in everyday life (Jackson 2002; Poria et al. 2017; Van Sterkenburg 2003).

Given the increasingly complex needs of contemporary language users — particularly due to the dominance of virtual communication channels — modern sustainable lexicography should provide additional insights into the semantic nuances that determine the emotional impact of a lexeme on both the speaker and the interlocutor. In this light, dictionaries should serve not only as tools for communication but also as instruments for shaping the emotional and cultural norms of a language community.

The process of assigning meaning and definitions to specific lexical units should be preceded by an in-depth analysis of lexicographic choices made by native speakers of the language. This includes not only how certain terms have developed through historical and cultural interactions within specific usage patterns, as emphasized by Hanks (2013), but also a more profound understanding of their emotional effects on interlocutors in modern usage. One such analysis is offered in this paper. Thus, contemporary lexicography should contribute to understanding how specific lexemes elicit positive, negative, or neutral reactions among speakers of different languages. In this way, dictionary users — both non-native speakers and those seeking more appropriate expressions in various communicative contexts — would receive comprehensive information about the meaning and emotional weight of the lexemes they intend to use.

As suggested by Lazarus (1991), emotions are directly and inseparably linked to cognitive evaluations of situations. Words, as carriers of meaning, therefore have the potential to act as triggers for different emotional responses. These insights are further developed within the framework of modern digital communication, which is predominantly written and highlights the need for a new perspective on lexicography — one that adapts to the requirements of contemporary speakers.

Currently available dictionaries for Serbian and Montenegrin speakers tend to treat lexemes at a general and basic contextual level. We believe it is important

to explain the contexts in which certain words are used, the communicative intentions behind them, and the ways in which they can alter the emotional dynamics of communication. It is well known that a word can have a neutral meaning in one context but take on strong emotional significance in another, and that very few lexemes in dictionaries are described with only a single definition. Typically, such words belong to invariable parts of speech¹.

Contemporary technologies have created ideal conditions for making lexicography sustainable by enriching it with interactive elements such as audio-visual content, which would help speakers find the most appropriate terms for a given situation. To become sustainable, lexicography should expand through such interactive elements — for instance, audio-visual recordings that illustrate how words sound in different emotional tones and scenarios. This would enable dictionary users to better understand the emotional weight of words.

When designed in this way, a modern dictionary becomes a handbook containing recommendations for the application of specific lexical items, serving the needs of modern society — not only for those who are non-native speakers but also for native speakers uncertain about how to fine-tune their expressions, whether in everyday communication or in situations requiring particular verbal caution.

4. Modal verbs, modalized expressions, and emotional intelligence

The primary function of language is to serve as a medium through which human beings — as inherently communicative creatures — connect with others. This connection occurs through the reception or transmission of messages, either with or without conscious intent, and these messages are often processed emotionally by the receiver. The ability to regulate emotions by understanding how one's language affects others falls within the domain of emotional intelligence, a concept foundationally developed by Daniel Goleman (1995). Often considered more crucial than traditional intellectual competence, emotional intelligence is seen as essential to both the cognitive and communicative capacities of a language user (Mayer et al. 2008). Emotions are channeled through language, resulting in the formation of positive interpersonal relationships among speakers of a given language (Petrides and Furnham 2001).

Consequently, research into the emotions triggered by specific lexical items should be an integral part of contemporary lexicography. Such work would allow dictionaries to function as multifaceted linguistic handbooks, aligned with the fast-paced and emotionally nuanced demands of modern communication. Understanding the meaning of a modal verb and its full implications is inseparable from the context of emotional intelligence and the regulation of emotions — not only those of the speaker but also those of one or more interlocutors. Thus, modal verbs are seen as fundamentally significant in both linguistic and psychological analyses.

Perceived from this perspective, modal verbs are not only indicators of possibility, obligation, permission, or volition but also markers of emotional states. Their use reflects the speaker's ability to recognize and regulate emotions by selecting modal constructions with varying degrees of emotional intensity. Goleman (1995) defines emotional intelligence as the ability to identify, understand, and manage emotions. When considered alongside findings by Mayer et al. (2008) and Petrides and Furnham (2001), it becomes evident that a lexicographic approach to modal verbs must be a central component of sustainable, contemporary lexicography. In this sense, the intentional and strategic use of specific lexical items and expressions signifies a speaker's awareness of their own emotional states and their communicative impact on others.

Many linguists have studied the role of modal verbs in expressing beliefs, attitudes, and intentions, generally agreeing on their importance for conveying tonal nuance in specific contexts (Coates 1983; Goleman 1995; Hyland 1998; Palmer 2001). The speaker's decision to use a specific modal auxiliary not only conveys fine-grained semantic distinctions but also indirectly intensifies the emotional tone of the message, which is of great importance in shaping communicative strategies, whereby language can be rendered more or less polite (Trbojević-Milošević 2004; Milošević 2008; Mrdak-Mićović 2013, 2016). Hence, the phenomenon of *linguistic emotional load* is especially relevant and can be closely linked to the system of modal verbs. When combined with intonation and context, modals can enhance empathy or provoke resistance in the interlocutor (Bybee et al. 1994; Coates 1983; Lyons 1977; Palmer 2001). Since one of the fundamental tasks of lexicography is to map the dynamics of meaning, analyzing modal verbs can deepen our understanding of emotional tone in speech and support more effective management of social and professional interactions.

Modal verbs can, therefore, be seen as essential tools for mediating between the rational and emotional domains of human expression. They are particularly suitable for research of this kind, given the intrinsic link between language and the cultural and psychological aspects of human identity. Their multifaceted role in contemporary language — scientifically recognized — makes them versatile instruments: they may function as expressions of support and affirmation, or as imperatives, depending on the communicative context. Accordingly, a speaker can use them to convey a positive emotional tone or, conversely, to demand or prohibit something.

Understanding these subtleties is fundamental to the development of sustainable, modern lexicography. This discipline must be viewed not merely as an inseparable part of a language's cultural code and specificities but also as a valuable source of insight into the connotation, intonational profile, and affective dimension of not only modal verbs but lexical units more broadly. Such an approach would significantly contribute to a more effective understanding of the dynamics of emotion and communication among speakers of a given language, in today's globalized society — thereby fulfilling the fundamental mission of contemporary sustainable lexicography.

5. The modal verb *moći* in the *Dictionary of the Serbian Language and the Encyclopedic English–Serbian Dictionary*

The *Rečnik srpskog jezika* (Dictionary of the Serbian Language), published by Matica srpska (1994–2013), has been, widely, regarded by leading scholars (Katičić 1987; Klajn 2005; Mrkić 2010; Radovanović 2000) as the principal lexicographic foundation of contemporary Serbian. It also represents a fundamental source for linguistic research focused on the structure and development of the Serbian language. This dictionary is a monolingual normative reference work intended primarily for native speakers, as explicitly stated in its preface, which defines its purpose as providing "an authoritative linguistic resource for the educated speaker" (Matica srpska 1994). Non-native speakers are not identified as part of the primary target audience, although the dictionary can be — and indeed sometimes is — used for language learning purposes, especially by advanced learners with high levels of language proficiency.

The *Rečnik* offers a comprehensive overview of lexical items, including in-depth semantic analyses and examples of usage across various contexts. The editors aim to encompass both archaic and contemporary forms, thereby granting users insight into the diachronic evolution of the lexicon, as well as its regional and dialectal variation.

Nevertheless, the practical usability of this dictionary is significantly limited by its lack of digital accessibility. The absence of searchable databases and interactive digital tools notably reduces its effectiveness for today's users — especially non-native speakers — who typically require fast, context-sensitive access to linguistic information (Atkins and Rundell 2008). Moreover, the descriptions of lexical entries — as well as of grammatical structures and idiomatic expressions — are often linguistically dense and syntactically complex, which can render the material inaccessible to a broader readership, including learners of Serbian as a foreign language.

From a methodological perspective, there remains considerable room for the improvement of lexicographic practice, particularly through the adoption of corpus-based approaches. Such methods must be both rigorous and aligned with current trends in language use in order to ensure the continued relevance of the dictionary in contemporary linguistic inquiry. As Kilgarriff and Tugwell (2001) emphasize, "a combination of significant patterns may lead to even greater precision in the identification of collocations," underscoring the importance of a systematic, corpus-driven approach to dictionary compilation and lexical description.

Within the *Rečnik srpskog jezika*, the modal verb *moći* is presented as a polysemous lexical item, encompassing a wide range of semantic nuances across both simple and complex syntactic structures. The lexicographers address epistemic, deontic, and dynamic modalities, while also attempting to distinguish between so-called "literary" and "contemporary" usage. It is important to clarify here that the term "literary" refers to the standardized, codified form of the language, rather than to the language of literary texts in the artistic sense (Kapetanović 2016).

МОТРИТИ имѣф. watch (на некога sb како неко чини нешто sb do sth; нешто sth: so that nothing bad happens)

МОЋ, **МОЋИ**(-) ж 1. might (јачина); power (способност) ability, телесна, умна of body/mind); powers (pl; расуђивања, зајажња, убеђивања of reasoning/observation/conviction); vigor AE, vigour BE (крејкост); potency (лека); без -и powerless; (нешто) биши ван нечије -и (sth) be beyond one's control; имаши м. да се учини нешто have the p. to do sth; м. навике force of habit; м. памћења memory; раслолажашти моћи wield p.; средње -и (жен.) medium-yield; учиниши све у својој моћи do one's best 2. hold (нај појединцем); sway (власт); тусете (држљивена); validity (законска); даши законску м. Нечему enact sth; даши некоме м. (овластити) give sb authority/power; добиши м. gain power; имаши м. наг have a h. on/over; имаши некога у својој -и have a hold on one's mind

МОЋАН, **-АН** -ириг. 1. mighty; powerful (особа, зрупа) 2. potent (лек, оружје)

МОЋИ, **ПРЕЗ. МОЋУ** 1. л. јд., можеш 2. л. јд., може 3. л. мн. имѣф. 1. can (pt could; физички physically, умно mentally); get to do (нешто што се одавно желело); Без ње/њега се не може. She/He is indispensable; м. више (јести) have had enough; Не моју више. (од умора) I am knocked up; Не може се знаши... There is no knowing...; не м. а да се не учини нешто cannot help doing sth; нико не м. can't possibly (учиниши нешто do sth); Све се може кад се хоће. When there's a will there's a way; чиниши колико/што се може do one's best; Шта ја шу моју. I can't help it; 2. can (pt could), fort may (дозвола permission); Моју ли? С. П? Моју ли Вас нешто ишиши? Might I ask a question? fort; Може ли (кол.) Је л' може (јои) вина? Won't you have some (more) wine?; Може му/јој се (да чини нешто) He/She is in a position to do it/sth; Може му/јој се да зрешу. He/She gets away with murder. infml; м. без do without, с. manage without; м. издржашти без нечега, м. забораваши на нешто (јер више није потребно) с. dispense with sth; Не може јој/му се веровати. She/He is not to be believed; Не може јо јоко; Тако не може биши. That/This won't do; не м. сапот; не м. без can't do without; Нему/Њој се све може. He/She can get away with anything; Рекла је да моју да позајмам. She said I might borrow it; У њега/њу се не може(и) поуздаши. He/She is not to be relied on. 3. can (pt could), may (pt might) (пошијање моуности possibility); би могао (вероватно) could; Кага ћу га м. наћи? When will he be available?; могао је лако п. easily/well (+ have + pp); може лако т. easily/well (+ have + pp); Могао би доћи ако буде лепо време. He might come if the weather is nice; Могао би се обогашиши да је вредно радио. He might have become rich if he had worked hard; Мола би ишиши. She might leave; Мола је узешти такси. (али није) She could have taken a taxi; Моли бисмо (кад нема ништа боље)... We m./might as well...; Моли бисмо оставиши на столу. You might leave it on the table. fort; Може лако обо-

леши. He/She is apt to fall ill; може само да користи некоме sb cannot help but benefit; м. ишасти и овако и онако could go either way; не може биши (sth) can't be; није могао/мола биши couldn't have been; Нисам могао позрешити. I can't have made a mistake; сасвим лако м. учиниши нешто (знајући нега/њу) be quite capable of doing sth; Шта ли се све мола десити? What с. have happened? 4. (очекивање; be expected to) Како можеш да шо радиши? (То није у реду.) How can you find it in your heart to do that?; Могао би да се ошиши. You could do with a haircut; Могао би ми однеши ово писмо на ишиши. You might post this letter for me.; То се [и] мола очекиваши. That was likely to happen. 5. Може! (Одговара ми) (Sth) is good!; It's all right!; Све може. (ма како неприхватљиво/необично) Anything goes.

МОЊИК м mogul /'mægʌl/; -ици (мн.) the powers that be infml

МОЊНО ирил. powerfully, forcefully; -ци powerfulness, mightiness

МОУЗЛИ м: Хенри Гвајн-Цефрејс М. (енглески фризичар) Henry Gwyn Jeffreys Moseley

МОФЕТА ж mofette (испарење уљендиоксида)

МОХАЧКИ ириг. the Battle of Mohács (a town in southern Hungary; in 1526, where the Turks defeated the Hungarians and the Serbs; in 1687 when the Germans defeated the Turks)

МОХЕР м mohair

МОХИКАНАЦ, **-НИ** м Mohegan, Mohican; „Последњи М.“ (роман Џемса Фенимора Купера) The Last of the Mohicans

МОХО м moho (раван у Земљиној кори)

МОХОРОВНИЧЕВ ириг.: -а раван дисконцинишиши Mohorovičić discontinuity

МОЦАРЕЛА ж mozzarella (сир)

МОЦАРТ м Mozart /'mɔʊtsɑ:t/

МОЧА ж piece of bread dipped in gravy

МОЧАЛИВА ж дијал. dub (dial; мочвара)

МОЧАР м књиж. damp (влага у ирироди)

МОЧАРАН, **-РН** -ириг. књиж. damp (in nature)

МОЧВАРА ж marsh, swamp, morass

МОЧВАРАН, **-РН** -ириг. 1. swampy (ливада meadow); -рно шло/землишиши marsh incount 2. -рна зронцица swamp fever

МОЧВАРИЦА ж wader

МОЧВАРИЦА ж заси. malaria

МОЧВАРНОСТ ж swampiness, marshiness

МОЧВАРУША ж в. мочварица

МОЧЕЊЕ с soaking; м. освеженог хлеба у вину intinction

МОЧИЛО с mordant

МОЧИЛСКИ ириг. mordant (боја dye)

МОЧИТИ имѣф. 1. soak (квасити); dip (умакаши);

dunk (ири јелу; food у in, into) 2. м. се soak, steep (vi)

МОЧИЦА ж small pole

МОЧУГА, **ДАИ**/лок. узм ж club (stick)

МОШЕЈА ж mosque /mɔsk/

МОШКОВИЦА ж moschatel /mɔskə'tel/ (биљка)

МОШНИЦЕ ж мн. scrotum (pl -tums/ta)

МОШНИЧНИ ириг. scrotal; а кеса scrotocele

Figure 2: The modal verb *moći* as described in the *Encyclopedic English–Serbian and Serbian–English Dictionary* (Hlebec 2003).

The Encyclopedic English–Serbian and Serbian–English Dictionary by Boris Hlebec (2003) represents one of the most significant bilingual dictionaries in the region. According to the author's preface, its intended audience includes both native speakers of Serbian learning English and English-speaking learners of Serbian. In terms of its scope and didactic intent, it may be reasonably compared to the *Oxford Advanced Learner's Dictionary* (Hornby et al. 2015).

Although the verb *moći* is treated in considerable semantic and stylistic detail, Hlebec's approach is grounded in traditional descriptive linguistics, without the incorporation of more recent pragmatic or corpus-informed methodologies, which, as Béjoint (2010) observes, are increasingly expected in contemporary lexicographic practice. This limits the dictionary's ability to reflect authentic usage patterns and real-world communicative contexts.

Modern dictionary users increasingly expect lexicographic descriptions to be rooted in authentic discourse and to reflect real-life language usage, especially in the domain of collocations and pragmatic connotations (Hanks 2013). In this respect, Hlebec's treatment of idiomatic expressions frequently lacks consistency with current usage in either English or Serbian, which may lead to misinterpretation and reduced pedagogical value. Thus, the modernization of dictionaries that seek academic and educational relevance is not merely desirable but essential. The integration of corpus-based, multimedia-enhanced, and emotionally and pragmatically contextualized lexical descriptions constitutes a critical imperative for contemporary, sustainable lexicography (Béjoint 2010; Hanks 2013).

Contemporary native and non-native speakers alike increasingly require lexical resources that offer detailed, pragmatically oriented explanations — resources that function as practical guides across diverse communicative settings. Corpus-driven studies have confirmed that language users seek information not only about meaning and form but also about functional-pragmatic usage, including stylistic coloring and emotional resonance of expressions (Atkins and Rundell 2008; Hanks 2013; Tognini-Bonelli 2001). In addition, the inclusion of multimedia content — such as audio recordings and visual examples — could significantly enhance user comprehension and application of target lexical items. Of particular importance is the need to indicate contextual appropriateness, the emotional impact certain modal verbs may have on interlocutors, and the types of emotional cues that may trigger their use in discourse.

In the broader context of contemporary lexicography, the findings of Kilgarriff and Tugwell (2001) regarding the systematic identification of collocations via corpus analysis may be viewed as emblematic of a broader trend: the need for lexicographic descriptions to be grounded in empirical, statistically-supported analysis of actual language use. While their study is focused on specific computational tools such as Word Sketch, its implications extend to the general methodological framework of modern lexicography, which strives for greater semantic precision, pragmatic relevance, and contextual richness. Such demands are particularly evident in relation to the needs of today's language users — both native and non-native — who seek nuanced, usage-sensitive lexical guidance. This is consistently demonstrated in corpus-based research (Atkins and

Rundell 2008; Tognini-Bonelli 2001; Hanks 2013), which underscores the importance of descriptive accuracy, frequency data, and situational appropriateness. Moreover, the integration of multimodal elements — including sound and image — has become an indispensable feature of modern dictionary resources. These additions support a multilayered understanding of word meaning and function. Special attention should be given to the pragmatic functions and emotional valence of expressions — particularly modal verbs, which not only reflect the speaker's stance but also shape the interpretation of meaning in interaction. In this light, pilot studies offer valuable insights into the future development of digitally-oriented, corpus-supported lexicographic practice, regardless of language or target audience. This is precisely where the present study contributes. The modal verb *moći* proves to be especially representative and lexically transparent, as it captures with unique sensitivity a range of meanings associated with possibility, permission, ability, and subjective disposition. In this sense, *moći* serves as an ideal lens for examining the conceptual layering of language through the prism of emotional and pragmatic competence. By analyzing its frequency, syntagmatic behavior, and functional distribution, we can most clearly illustrate a central proposition of this study: that emotional intelligence is inseparable from modern lexicographic reflection, and that it must be recognized as a key category in the interpretation of meaning within the digital linguistic environment.

6. Research

The research was designed to identify the degree of emotional intensity and perception of the modal verb *moći* concerning the parameters of gender, age group, and place of residence among speakers from four distinct age groups in urbanized and peripheral areas of Montenegro.

A total of 684 respondents participated in the study, with the sample stratified according to key sociodemographic parameters. Gender representation was approximately equal, and the age groups were delineated as follows:

- Children and adolescents (12–18 years): 163 respondents²
- Adults (19–40 years): 176 respondents
- Older adults (41–65 years): 170 respondents
- Seniors (66+ years): 175 respondents

Age is a particularly salient variable in sociolinguistic and lexicographic research, as it correlates with language change, lexical innovation, and varying degrees of linguistic conservatism (Eckert 1997; Trudgill 1986). Tracking responses across age cohorts enables the identification of diachronic shifts in verb perception and modality usage, which are often generationally mediated. Equally important is the inclusion of participants' place of residence. Of the total sample, 348 respondents were from urban centers (e.g., Podgorica, Nikšić, Herceg Novi), while 336 resided in rural communities. Urban-rural distinctions are known to correspond

with variation in lexical choices, linguistic prestige norms, and exposure to standardized versus dialectal forms (Milroy and Milroy 1985; Chambers and Trudgill 1998). The careful balance between urban and rural dwellers in the present study strengthens the ecological validity of the findings and supports more nuanced sociolinguistic inferences regarding the usage and perception of *moći*. However, it would be both insightful and methodologically sound to incorporate data on the educational background of participants, therefore, it may, nonetheless, contribute meaningfully to broader, multi-dimensional investigations of lexical perception. The level of formal education has been shown to significantly affect language awareness, metalinguistic ability, and lexical sensitivity (Labov 2006; Biber and Finegan 1994). Including this variable would enhance the interpretative depth of the study and allow for a more granular analysis of perceptual variation in relation to the verb *moći*. Education not only shapes language production and comprehension but also influences attitudes toward linguistic norms and usage, particularly in sociolinguistically sensitive domains. Furthermore, the integration of educational attainment, age distribution, and dwelling type constitutes a methodologically sound framework for interpreting variation in lexical perception. This triangulated demographic approach reflects best practices in contemporary sociolinguistic research, where intersecting social variables are recognized as co-constructors of linguistic behavior (Tagliamonte 2006; Chambers 2009). In lexicographic studies, such multidimensional profiling enables more accurate mapping of lexical variation and semantic interpretation across speaker populations (Atkins and Rundell 2008; Svensén 2009). These parameters — education, age, and place of residence — have been repeatedly shown to influence language usage patterns, lexical frequency, and the social indexing of meaning (Labov 2001; Aitchison 2012), and their combined inclusion enhances both the validity and interpretability of the findings. The study employed a mixed-methods approach, incorporating both structured questionnaires and in-depth interviews to generate comprehensive quantitative and qualitative data.

The **in-depth interviews**, conducted in **Montenegrin**, were designed to supplement the structured questionnaire by eliciting **richer, context-sensitive interpretations** of the verb *moći*, particularly in emotionally charged or socially complex scenarios. The goal was to explore participants' **subjective emotional responses** as well as **pragmatic interpretations** beyond the limitations of fixed-response items. The interviews employed a **semi-structured format**, a method widely recognized in qualitative sociolinguistic research for its flexibility and depth (Kvale 1996; Brinkmann and Kvale 2015). While the interviews were guided by a **set of predetermined open-ended questions**, the format allowed for follow-up prompts and elaboration depending on each respondent's responses. This method enabled the researcher to probe into **nuanced emotional and contextual associations** with the modal verb *moći*, especially regarding social hierarchies and speaker-hearer relationships.

Examples of the open-ended questions included:

- *What emotions does the verb 'moći' evoke when it implies obligation or permission?*

-
- *How do your emotions change depending on who uses the verb — a friend, stranger, or authority figure?*
 - *How do you interpret the phrase 'It could be better' during difficult times — as comforting or frustrating?*
 - *When you hear 'Can you finish this?' do you experience pressure, politeness, or neutrality?*

The respondents **answered orally in the presence of the researcher**, who **recorded all sessions** using digital audio equipment to ensure accuracy and reduce the potential for interpretive bias during note-taking. These recordings were later **fully transcribed** verbatim for qualitative analysis. This procedure aligns with standard best practices in linguistic fieldwork and discourse analysis (Duranti 1997; Johnstone 2000). The language of all interviews was **Montenegrin**, maintaining consistency with the language used in the questionnaire and ensuring that participants could express their emotions and interpretations in their native or most comfortable linguistic register. This linguistic consistency is critical for maintaining emotional validity and preventing translation-related semantic distortions, especially in studies focusing on emotionally loaded lexical items (Pavlenko 2005).

The transcribed interviews were subjected to **thematic coding**, using a **grounded theory approach** (Glaser and Strauss 1967; Charmaz 2006) to identify **recurring emotional and pragmatic themes**. Codes were derived inductively from the data and subsequently grouped into broader categories (e.g., social distance, modality intensity, emotional ambiguity). This allowed for systematic comparison across demographic groups and linkage to the quantitative findings. The interview data played a crucial role in **triangulating** the quantitative results, deepening insights into statistically significant emotional patterns that emerged from the questionnaire analysis. For instance, where ANOVA tests revealed generational differences in the perception of *moći* as empowering or pressuring, interview data contextualized these findings by illustrating how seniors associated the verb with obligation under authority, while younger respondents often linked it to agency and possibility.

7. Discussion

Aiming to assess the sociolinguistic dynamics underlying the perception of the modal verb *moći*, the analysis considers both demographic variables and age-based cohort distinctions. The following section presents quantitative findings that illuminate how emotional associations with *moći* vary systematically across generational groups and residential contexts, revealing patterns shaped by psychosocial and cognitive dimensions of language processing.

Quantitative findings indicate that **positive emotional associations with *moći* are strongest among younger urban participants**, particularly adolescents, whereas these associations diminish with age and rural residency. Conversely, **nega-**

tive connotations increase notably among older rural participants, likely reflecting decreased agency and autonomy, consistent with Charles and Carstensen (2010). The pattern also supports Đorđević et al. (2018) and Isaacowitz and Blanchard-Fields (2012) who emphasized sociopsychological influences on language processing.

Table 1: Average ratings of the emotional experience of the verb "to be able to" by gender, age group, and place of residence

Group	Positive Emotions (1–5)	Neutral Feelings (1–5)	Negative Emotions (1–5)
Children/ Adolescents (12–18) — Urban	4.2	3.1	2.5
Children/ Adolescents (12–18) — Rural	3.8	3.3	2.8
Adults (19–40) — Urban	4.0	3.0	2.7
Adults (19–40) — Rural	3.9	3.2	2.9
Older Adults (41–65) — Urban	3.6	3.4	3.1
Older Adults (41–65) — Rural	3.4	3.5	3.2
Seniors (66+) — Urban	3.3	3.6	3.5
Seniors (66+) — Rural	3.2	3.7	3.7

The table presents the average ratings of the emotional experience associated with the verb *moći* based on gender, age group, and place of residence, highlighting significant variations across groups. In general, positive emotions are more pronounced among younger participants, with children and adolescents from urban areas providing the highest ratings (4.2). As age increases, the intensity of positive affective responses decreases, particularly among seniors in rural areas (3.2). Neutral feelings show a slight tendency to increase with age, which may suggest an increased cognitive divergence from the meaning of the verb in line with previous studies indicating changes in emotional interpretation and experience of this modal verb over the years. Negative emotions are most pronounced among older groups, with seniors in rural areas reporting the highest levels of negative emotional experience (3.7). This finding aligns with research indicating a decline in the perception of control and autonomy in later life, as well as changes in emotional regulation and interactions during the later stages of life (Charles and Carstensen 2010). These results point to the complex interaction between age, place of residence, and emotional experience of modal constructions, which may have implications for linguistic and psychological models of language interpretation in different demographic groups. This finding is supported by the work of Đorđević et al. (2018) and Isaacowitz and Blanchard-Fields (2012), who emphasize the importance of understanding emotions and language in relation to sociological and psychological variables.

The thematic analysis of interview data reveals generational variation in emotional conceptualization of *moći*. For children and adolescents, the verb is primarily linked with **freedom, choice, and personal growth** — consistent with the tenets of Self-Determination Theory (Ryan and Deci 2000), which underscore autonomy and competence as key motivational factors. Among adults, urban participants interpret *moći* as representing **opportunity and active decision-making**, often in professional and social contexts. This underscores the perceived **autonomy embedded in urban life**, aligning with Inglehart and Welzel's (2005) view of modern societies as participation-oriented. In contrast, rural adults more frequently associate *moći* with **social norms and obligation**, highlighting how community structures and spatial context shape linguistic interpretation.

The senior respondents overwhelmingly perceive *moći* in terms of **constraint and limitation**, often attributing this to **declining health or social influence**. This reflects Baltes and Baltes' (1990) model of selective optimization with compensation, as well as Lachman and Weaver (1998), who associate diminished control with lower emotional well-being. Therefore, the responses from the thematic questionnaire indicate generational differences in the emotional and conceptual experience of the verb "to be able to," reflecting deeper sociocultural and psychological variations in how participants perceive this modal verb. Among children and adolescents, this verb is predominantly associated with freedom, opportunities, and possibilities, which points to the developmental aspect of personal autonomy and exploration of the world. According to Ryan and Deci's (2000) Self-Determination Theory, the perception of power among young individuals is determined by the level to which their basic psychological needs — autonomy, competence, and relatedness — are satisfied or upset in a particular social environment. It is both encouraging and expected that this age group perceives the verb through a positive lens, focused on the opportunities available to them in the contexts of education, social interactions, and independent decision-making.

Regarding adult participants, the research findings suggest that differentiation is observed depending on the participants' place of residence. Similar to younger participants, adults predominantly associate the verb "to be able to" regarding the concept of decision-making, which can be understood as the perception of a broader spectrum of opportunities offered by urban living, encompassing social, professional, educational, and, broadly speaking, the quality of life in modern society. Conversely, participants from rural areas associated the verb in accordance with the idea of permission or obligation, suggesting that individual autonomy is more frequently linked to social norms and collective responsibilities. This aligns with research on the differences in the perception of freedom depending on the social and spatial context (Inglehart and Welzel 2005). According to these authors, the concept of power is dynamic, and in modern, developed societies, it is more closely linked to the idea of participation rather than control. Therefore, the results from this part of the study reveal a clear connection to the findings of Ryan and Deci (2000), as well as Inglehart and Welzel (2005), since the social context is considered the most significant factor

in shaping or undermining common-sense thinking, a sense of security, and motivation.

As anticipated, taking into account previous theories and the findings of Baltes and Baltes (1990), who define the *model of selective optimization with compensation*, which also reflects a sense of reduced control over life and abilities, the oldest group of participants predominantly associated this modal verb with limitations, often related to health issues. These findings align with those of Lachman and Weaver (1998), who argue that the reduced perception of control among seniors regarding their own lives directly affects their decreased sense of well-being and emotional regulation in this age group.

The modal verb *moći* ("to be able to" and "to have to") is used in this article as a representative model to illustrate the imperative of integrating emotional intelligence as a significant parameter in contemporary lexicographic practice and in the semantic framing of lexemes. Through an analysis of its treatment in dictionaries still widely used within the Montenegrin linguistic context, this study underscores the necessity of incorporating emotional and pragmatic markers into modern lexicographic methodologies. This would be a significant step toward making contemporary lexicography sustainable and aligned with the needs and expectations of both native and non-native speakers. Our research, which investigates the emotional impact of these verbs on interlocutors and the emotions that trigger their usage, suggests that emotional annotations should be integrated into modern dictionaries, a step toward improving both their relevance and practical usage. Therefore, each lexical unit should be examined in the manner we proposed, and after thorough research, it should be marked accordingly, as emotional perception can strongly influence both the speaker and the recipient. This marking could be achieved through the use of color, indicating the average emotional intensity based on research findings. We suggest employing a traffic light method, utilizing three basic colors to denote the emotional intensity of a particular lexeme, thus providing a warning or encouragement for its use in a specific context. The use of visual symbols, such as emoticons, to represent emotional connotation would facilitate intuitive recognition of the dominant emotions associated with a specific lexical item. In addition, numerical scales of emotional intensity could be applied to indicate the average emotional strength of a lexeme or expression, based on research, or alternatively, comments could be offered within definitions.

Furthermore, the contextual notes should inevitably accompany those related to age groups, specifically the age of the speaker or interlocutor to whom a certain form is addressed, as this would provide additional clarification when defining particular expressions. Given that emotions in academic sources are recognized as fundamental not only in the experience of the world but also in its description, the accurately defined connotations of lexemes would significantly contribute to better pragmatic and intercultural language competence, which is the core idea behind contemporary sustainable lexicography. Therefore, we believe that modern lexicography should not be merely a descriptive linguistic discipline. To remain relevant and sustainable, it must integrate emo-

tional factors that influence language use. In this way, dictionaries become important emotional and cognitive tools for both native and non-native speakers, helping them to better understand each other.

Table 2: A tabular overview of possible approaches to defining the meaning of the verb 'moći'

Meaning / Context	Description	Strength	Color	Emojis	Stars	Icon	Font Style
1. Physical ability — "I can play the piano."	Has the ability to perform an action	Medium	Yellow		***		Medium
2. Permission — "Can I close the door?"	Has permission or authorization	Medium	Yellow		***		Medium
3. Possibility / Potential — "That can be true."	Something is possible, not certain	Weak	Blue		**		Small
4. Probability — "He might be at home now."	There is a chance, but it's uncertain	Weak	Blue		*		Small, italic
5. Polite request — "Can you help me?"	Not strong, rather polite and soft	Medium	Yellow		**		Medium, italic
6. Sarcasm / Irony — "Yeah, right, you can go out!"	Used ironically, meaning the opposite	Strong	Red		*** *		Large, italic
7. Total power — "He can achieve anything."	Extreme ability, dominance, full control	Strong	Red		*** **		LARGE BOLD
8. Inability (Negation) — "I can't believe you any more!"	Complete inability, often emotional reaction	Strong	Red		*** **		Bold, red

8. Conclusions

Artificial intelligence has long been taking over many areas of human activity, and one of an eminent field of scholarly investigation, in which its influence is felt is the area of education, learning, and foreign language teaching. Therefore, modern lexicography must keep up with contemporary trends and the evolving needs of individuals in modern contexts, who are accustomed to receiving information quickly and accurately. The establishment of contemporary online dictionaries requires thorough work, but once they are finalized, they will be of real benefit to speakers worldwide. Lexicography developed in this way will be sustainable, used not merely as a repository of linguistic information, but also as a dynamic and inclusive tool that reflects cultural diversity, and fosters global

understanding — making it a truly irreplaceable resource for humanity in the fullest sense. Therefore the aim of the present research is to serve as a pilot study, or more precisely, a case study focusing on one highly frequent modal verb present in both language systems. By examining this lexical item in detail, the study seeks to illustrate a possible approach to how such units can be represented in modern lexicographic resources. The intention is for this case to exemplify a broader methodology — one that we suggest could meet the evolving needs of dictionary users by offering more contextually and functionally informed entries.

9. Endnotes

1. **Invariable parts of speech** are lexical categories whose members do not undergo inflectional changes. Unlike variable word classes such as nouns or verbs, invariable categories (e.g., prepositions, conjunctions, adverbs, and interjections) maintain a fixed morphological form regardless of grammatical context. Their syntactic and semantic roles are determined by position and usage within the sentence, rather than by morphological variation (Huddleston and Pullum 2002).
2. Authors' note: The term *youngest cohort* refers specifically to participants aged **12 to 14**, who were distinguished from the broader 12–18 age group based on developmental considerations. Age was indeed used as a criterion, informed by research in sociolinguistics and educational psychology that emphasizes significant differences in linguistic awareness and cognitive processing between early adolescents (12–14) and older teenagers (15–18) (cf. Vygotsky 1978; Eckert 2000). Accordingly, the adapted version of the questionnaire — with simplified phrasing and reduced metalinguistic load — was administered **only to the 12–14 subgroup**, ensuring that the instrument remained cognitively and linguistically appropriate for their developmental stage. Participants aged 15–18 received the standard version, as preliminary piloting showed that they were capable of engaging with the full set of items without confusion or fatigue.

10. References

- Aitchison, J. 2012. *Words in the Mind: An Introduction to the Mental Lexicon*. Fourth edition. Oxford: Wiley-Blackwell.
- Apresjan, J.D. 2010. *Prospekt akcionogo slovarja ruskogo jazyka*. Moskva: Jazyki slavjanskih kultur.
- Atkins, B.T.S. and M. Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford: Oxford University Press.
- Bartsch, R. 1987. *Norms of Language: Theoretical and Practical Aspects*. London/New York: Longman.
- Béjoint, H. 2010. *The Lexicography of English: From Origins to Present*. Oxford/New York: Oxford University Press.
- Biber, D. and E. Finegan (Eds.). 1994. *Sociolinguistic Perspectives on Register*. Oxford/New York: Oxford University Press.
- Brinkmann, S. and S. Kvale. 2015. *InterViews: Learning the Craft of Qualitative Research Interviewing*. Third edition. Thousand Oaks, CA: SAGE.

- Bybee, J., R. Perkins and W. Pagliuca.** 1994. *The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World*. Chicago: University of Chicago Press.
- Calvo, R.A. and S. D'Mello.** 2010. Affect Detection: An Interdisciplinary Review of Models, Methods, and Their Applications. *IEEE Transactions on Affective Computing* 1(1): 18-37.
- Chambers, J.K.** 2009. *Sociolinguistic Theory: Linguistic Variation and Its Social Significance*. Second revised edition. Chichester, UK/Malden, MA: Wiley-Blackwell.
- Chambers, J.K. and P. Trudgill.** 1998. *Dialectology*. Cambridge: Cambridge University Press.
- Charles, S.T. and L.L. Carstensen.** 2010. Social and Emotional Aging. *Annual Review of Psychology* 61: 383-409.
<https://doi.org/10.1146/annurev.psych.093008.100448>
- Charmaz, K.** 2006. *Constructing Grounded Theory: A Practical Guide through Qualitative Analysis*. London/Thousand Oaks, California: SAGE.
- Chomsky, N., I. Roberts and J. Watumull.** 2023. The False Promise of ChatGPT. *The New York Times*, 8 March 2023.
<https://www.nytimes.com/2023/03/08/opinion/noam-chomsky-chatgpt-ai.html>
- Coates, J.** 1983. *The Semantics of the Modal Auxiliaries*. London: Croom Helm.
- Colman, L.** 2016. Sustainable Lexicography: Where To Go from Here with the ANW (*Algemeen Nederlands Woordenboek*, an Online General Language Dictionary of Contemporary Dutch)? *International Journal of Lexicography* 29(2): 139-155.
<https://doi.org/10.1093/ijl/ecw008>
- De Schryver, G.-M.** 2023. Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography* 36(4): 355-387.
<https://doi.org/10.1093/ijl/ecad021>
- Đorđević, D., et al.** 2018. Narativizacija društvenih događaja u porodici: Analiza životnih prekretnica kod pripadnika tri generacije. *Kvalitativna istraživanja u društvenim naukama: Od ličnog iskustva do socijalnih praksi*. Beograd: Filozofski fakultet.
- Duranti, A.** 1997. *Linguistic Anthropology*. Cambridge/New York Cambridge University Press.
- Eckert, P.** 1997. Age as a Sociolinguistic Variable. Coulmas, F. 1997. *The Handbook of Sociolinguistics*: 151-167. Oxford: Blackwell.
- Eckert, P.** 2000. *Linguistic Variation as Social Practice: The Linguistic Construction of Identity in Belten High*. Oxford: Blackwell.
- Fuertes-Olivera, P.A.** 2024. Making Lexicography Sustainable: Using ChatGPT and Reusing Data for Lexicographic Purposes. *Lexikos* 34(1): 123-140.
<https://doi.org/10.5788/34-1-1883>
- Geeraerts, D.** 2010. *Theories of Lexical Semantics*. Oxford: Oxford University Press.
- Glaser, B. and A. Strauss.** 1967. *The Discovery of Grounded Theory: Strategies for Qualitative Research*. New Brunswick/London: Aldine Transaction.
- Goleman, D.** 1995. *Emotional Intelligence: Why It Can Matter More than IQ*. New York: Bantam Books.
- Hall, E.T.** 1976. *Beyond Culture*. Garden City, New York: Anchor/Doubleday.
- Hanks, P.** 2013. *Lexical Analysis: Norms and Exploitations*. Cambridge, Mass./London: MIT Press.
- Hlebec, B.** 2003. *Enciklopedijski englesko-srpski i srpsko-engleski rečnik*. Beograd: Zavod za udžbenike i nastavna sredstva.
- Hornby, A.S., M. Ashby and S. Wehmeier.** 2015. *Oxford Advanced Learner's Dictionary*. Ninth edition. Oxford: Oxford University Press.

-
- Huddleston, R. and G.K. Pullum.** 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.
- Hyland, K.** 1998. *Hedging in Scientific Research Articles*. Amsterdam/Philadelphia: John Benjamins.
- Inglehart, R. and C. Welzel.** 2005. *Modernization, Cultural Change, and Democracy: The Human Development Sequence*. Cambridge: New York: Cambridge University Press.
<https://doi.org/10.1017/CBO9780511790881>
- Isaacowitz, D.M. and F. Blanchard-Fields.** 2012. Linking Process and Outcome in the Study of Emotion and Aging. *Perspectives on Psychological Science* 7(1): 3-17.
- Jackson, Howard.** 2002. *Lexicography: An Introduction*. London/New York: Routledge.
- Johnstone, B.** 2000. *Qualitative Methods in Sociolinguistics*. Oxford/New York: Oxford University Press.
- Kapetanović, A.** 2016. *Standardni jezik i književna norma*. Zagreb: Institut za jezik.
- Kilgarriff, A. and D. Tugwell.** 2001. WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography. *Workshop Proceedings. COLLOCATION: Computational Extraction, Analysis and Exploitation (39th Annual Meeting of the Association for Computational Linguistics and 10th Conference of the European Chapter of the Association for Computational Linguistics, Toulouse, 7 July 2001)*: 32-38.
- Kvale, S.** 1996. *InterViews: An Introduction to Qualitative Research Interviewing*. London/Thousand Oaks, California: SAGE.
- Labov, W.** 2001. *Principles of Linguistic Change. Volume 2: Social Factors*. Malden, MA: Blackwell.
- Labov, W.** 2006. *The Social Stratification of English in New York City*. Second edition. Cambridge: Cambridge University Press.
- Lachman, M.E. and S.L. Weaver.** 1998. The Sense of Control as a Moderator of Social Class Differences in Health and Well-being. *Journal of Personality and Social Psychology* 74(3): 763-773.
<https://doi.org/10.1037/0022-3514.74.3.763>
- Lazarus, R.S.** 1991. *Emotion and Adaptation*. Oxford: Oxford University Press.
- Lyons, J.** 1977. *Semantics. Volume 2*. Cambridge: Cambridge University Press.
- Matica srpska.** (1994–2013). *Rečnik srpskog jezika (Vols. 1–12)*. Novi Sad: Matica srpska.
- Mayer, J.D., P. Salovey and D.R. Caruso.** 2008. Emotional Intelligence: New Ability or Eclectic Traits? *American Psychologist* 63(6): 503-517.
- Mesquita, B., J. de Leersnyder and M. Boiger.** 2021. *The Emotional Dynamics of Acculturation*. Cambridge: Cambridge University Press.
- Milroy, J. and L. Milroy.** 1985. *Authority in Language: Investigating Standard English*. London/New York: Routledge & Kegan Paul.
- Mrdak-Mićović, M.** 2013. Sociolingvistički aspekti iskazivanja zahvaljivanja i čestitanja: Parametri regija i životna dob. *Zbornik Matice srpske za filologiju i lingvistiku* 59(1–2): 123-137.
- Mrdak-Mićović, M.** 2016. Forme učtivosti u okviru govornog čina molbe u odnosu na parametre regija i starost u crnogorskom jeziku. *Folia Linguistica et Litteraria* 5(1): 45-58.
- Palmer, F.R.** 2001. *Mood and Modality*. Second edition. Cambridge/New York: Cambridge University Press.
- Pavlenko, A.** 2005. *Emotions and Multilingualism*. Cambridge/New York: Cambridge University Press.
- Pavlenko, A.** 2014. *The Bilingual Mind: And What It Tells Us About Language and Thought*. Cambridge: Cambridge University Press.
- Petrides, K.V. and A. Furnham.** 2001. Trait Emotional Intelligence: Psychometric Investigation with Reference to Established Trait Taxonomies. *European Journal of Personality* 15(6): 425-448.
<https://doi.org/10.1002/per.416>

- Poria, S., E. Cambria, R. Bajpai and A. Hussain.** 2017. A Review of Affective Computing: From Unimodal Analysis to Multimodal Fusion. *Information Fusion* 37: 98-125.
- Ryan, R.M. and E.L. Deci.** 2000. Self-determination Theory and the Facilitation of Intrinsic Motivation, Social Development, and Well-being. *American Psychologist* 55(1): 68-78.
<https://doi.org/10.1037/0003-066X.55.1.68>
- Sinclair, J.** 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Svensén, B.** 2009. *A Handbook of Lexicography: The Theory and Practice of Dictionary-making*. Cambridge: Cambridge University Press.
- Šubarić, S. and J. Đurčević.** 2023. The Names of the Balkan Peoples and the Names of the Inhabitants of Balkan Countries in Lexicography (on the Example of the *Dictionary of Montenegrin National and Literary Language*). *Lexikos* 33(1): 68-89.
<https://doi.org/10.5788/33-1-1797>
- Tagliamonte, S.A.** 2006. *Analysing Sociolinguistic Variation*. Cambridge: Cambridge University Press.
- Tognini-Bonelli, E.** 2001. *Corpus Linguistics at Work*. Amsterdam/Philadelphia: John Benjamins.
- Topolinjska, Z.** 2002. Šta tražim u rečniku? *Deskriptivna leksikografija standardnog jezika i njene teorijske osnove*: 33-37. Beograd: Institut za srpski jezik SANU; Novi Sad: Matica srpska.
- Trbojević-Milošević, I.** 2004. *Modalnost, sud, iskaz: Epistemička modalnost u engleskom i srpskom jeziku*. Beograd: Čigoja Štampa.
- Trudgill, P.** 1986. *Dialects in Contact*. Oxford/New York: Blackwell.
- Tubbs, S.L.** 2012. *Human Communication: Principles and Contexts*. Thirteenth edition. New York: McGraw-Hill.
- Van Dijk, T.A.** 2008. *Discourse and Context: A Sociocognitive Approach*. Cambridge: Cambridge University Press
- Van Dijk, T.A.** 2009. *Society and Discourse: How Social Contexts Influence Text and Talk*. Cambridge/New York: Cambridge University Press.
- Van Sterkenburg, P. (Ed.).** 2003. *A Practical Guide to Lexicography*. Amsterdam/Philadelphia: John Benjamins.
- Vygotsky, L.S.** 1978. *Mind in Society: The Development of Higher Psychological Processes*. (Edited by Cole, M., V. John-Steiner, S. Scribner and E. Souberman). Cambridge, MA/London: Harvard University Press.
- Wierzbicka, A.** 1992. *Semantics, Culture, and Cognition: Universal Human Concepts in Culture-specific Configurations*. Oxford/New York: Oxford University Press.
- Wierzbicka, A.** 1999. *Emotions across Languages and Cultures: Diversity and Universals*. Cambridge: Cambridge University Press.

Making Dictionary Content Accessible for People with Visual Impairments

Geraint Paul Rees, *Translation and Language Sciences,
Pompeu Fabra University, Barcelona, Spain*
(geraintpaul.rees@upf.edu)
(<https://orcid.org/0000-0002-9204-8073>)

Abstract: This article discusses the development and evaluation of EDictViz: a visually accessible dictionary website intended to meet the accessibility needs of people with visual impairments. It is argued that dictionary access is important not only because of the utilitarian role that dictionaries play in resolving language queries but also because of their potential to promote social inclusion. With reference to previous research, a summary is provided of the barriers to access typically faced by people with visual impairments when using dictionary websites. Following this, the way in which these problems have been addressed in the development and evaluation process of EDictViz thus far is discussed. An evaluation of a development version of EDictViz based on the Web Content Accessibility Guidelines is reported — it indicates that the latest prototype version of the tool is accessible. This suggests that the design decisions taken at the outset of the project were sound. Throughout this process the parallels between lexicography research, which is often concerned with providing efficient access to lexicographic data, and accessibility research in general are drawn. Plans for future experimental studies to evaluate EDictViz which directly involve people with visual impairments are set out. This development process demonstrates the potential for mutually beneficial collaboration between accessibility and lexicography researchers.

Keywords: ACCESSIBILITY, ADVERTISEMENTS, EDUCATION, LANGUAGE LEARNING, LANGUAGE TEACHING, LEXICOGRAPHY, LEXICOTAINMENT, ONLINE DICTIONARIES, WCAG, WEB DEVELOPMENT

Opsomming: Toeganklike woordeboekinhoud vir mense met visuele gestremdhede. Hierdie artikel bespreek die ontwikkeling en evaluering van EDictViz: 'n visueel toeganklike woordeboekwebwerf wat bedoel is om aan die toeganklikheidsbehoefte van mense met gesiggestremdhede te voldoen. Daar word aangevoer dat woordeboektoegang belangrik is, nie net vanweë die utilitaristiese rol wat woordeboeke speel in die oplossing van taalnavrae nie, maar ook vanweë hul potensiaal om sosiale insluiting te bevorder. Met verwysing na vorige navorsing word 'n opsomming gegee van die hindernisse tot toegang wat mense met gesiggestremdhede tipies teëkom wanneer hulle woordeboekwebwerwe gebruik. Vervolgens word die manier bespreek waarop hierdie probleme tot dusver in die ontwikkelings- en evalueringsproses van EDictViz aangepak is. 'n Evaluering van 'n ontwikkelingsweergawe van EDictViz gebaseer op die Webinhoudtoeganklikheidsriglyne word aangebied — dit dui daarop dat die nuutste prototipe weergawe van die instrument toeganklik is. Die ontwerpbesluite wat aan die begin van die projek geneem is, is derhalwe sinvol. Gedurende hierdie

proses word die parallelle getrek tussen leksikografiese navorsing, wat dikwels gemoeid is met die verskaffing van doeltreffende toegang tot leksikografiese data, en toeganklikheidsnavorsing in die algemeen. Planne word uiteengesit om EDictViz te evalueer aan die hand van toekomstige eksperimentele studies wat mense met gesiggestremdhede direk betrek. Hierdie ontwikkelingsproses demonstreer die potensiaal vir wedersyds voordelige samewerking tussen toeganklikheids- en leksikografiese navorsers.

Sleutelwoorde: TOEGANKLIKHEID, ADVERTENSIES, ONDERWYS, TAALANLEER, TAALONDERRIG, LEKSIKOGRAFIE, LEKSIKOVERHOUDING, AANLYN WOORDEBOEKE, WCAG, WEBONTWIKKELING

1. Introduction

This article reports on the development and evaluation of EDictViz¹: an online visually pared-down dictionary interface intended to be accessible for people with visual impairments. Firstly, the project is contextualised in a brief discussion of the problems involved in presenting lexicographic data in a visually accessible manner in traditional paper-based resources. These are juxtaposed with the affordances offered by online dictionaries. Then, in Section 2, the findings of recent research on the accessibility of online dictionaries for people with visual impairments are summarised. In Section 3, EDictViz is posited as a potential solution to the accessibility problems encountered in previous research. In Section 3.1, its development and evaluation thus far are described, while in Section 3.2, plans for future evaluation studies are outlined. The article concludes with a summary of some of the lessons learned from the project thus far and some of its wider implications for accessibility studies and lexicography.

Dictionaries are important. From a utilitarian perspective, they are used to search for the meanings of words and phrases, examples of their use in context, synonyms and antonyms, spellings, pronunciation, and etymological information, along with many other utilitarian functions. Beyond these utilitarian functions, they have a less obvious socio-political role, namely "giving tangible testimony to the status or identity of a language-speaking community" (Lew 2015a: 1). For example, dictionaries of lesser spoken languages attempt to mark out the boundaries of these languages by excluding loan words from more widely spoken neighbouring languages. Moreover, there is a generalised respect or reverence towards dictionaries in society. For example, the idea that the dictionary is the ultimate arbiter of wordhood is common if unjustified. Similarly, the ability to use a dictionary is sometimes perceived as a mark of an educated person. From this perspective, dictionary access can be seen as an important component in participation in society. In lexicography, there has been growing recognition of the importance of making lexicographic resources accessible in recent years. This is evidenced by recent congresses of the European Association for Lexicography (EURALEX). For example, the theme of the XIX congress (in Alexandroupolis, Greece, 2021) was *Lexicography for Inclusion* (Gavriilidou et al. 2021)

while the XXI congress (in Cavtat, Croatia, 2024) was preceded by a workshop on *Lexicography and Accessibility* (Rees et al. 2024).

There is a vast literature on the affordances of online dictionaries relative to their paper counterparts. Lew (2015b) provides a summary of user research on the use of online dictionaries until 2015. A commonly cited advantage to online dictionaries is that they are not subject to the space constraints of paper resources. However, while the space constraints faced by online dictionary makers are less acute than those faced by the creators of paper-based resources, there is nonetheless a limit on the amount of information that can usefully be displayed on screen at any one time on a dictionary webpage (Lew 2011). The speed of access to the relevant entry and sense of a word provided by online dictionaries, compared with their paper counterparts, is a clearer advantage. For example, search functionality allows users to find the entry they need without having to flick through an alphabetical index. Features such as predictive text (Jackson 2018) and fuzzy matching for misspellings and partially remembered phrases can compensate for gaps in users' knowledge of a word or phrase's written form (Lew 2012), thereby making the look-up process more efficient. At the level of the entry for a particular word, the use of colour syntactic labels (e.g., noun, verb, adjective, adverb, preposition) permitted by online dictionaries has been shown to have benefits not only for look-up speed but also look-up accuracy and vocabulary retention (Dziemianko 2015).

The inclusion of multimedia elements permitted by the online format is also a key advantage (Rees 2025). In general, illustrations combined with verbal definitions have been shown to lead to greater vocabulary learning than illustrations or verbal definitions in isolation (Nesi 1998; Dziemianko 2022). The affordances of multimedia have been particularly valuable for sign language dictionaries (McKee and McKee 2013). For example, still images can be used to illustrate signs and search for them by handshape categories and the place of articulation. These are fundamental elements in the meaning of signs. Similarly, video allows the capture of the dynamic motion and changes in handshapes needed to interpret and articulate signs. The option to replay and watch in slow motion is obviously advantageous for those learning sign language.

All these affordances improve the user friendliness of dictionaries or their accessibility in the general sense of the word. The application of video and images in sign language dictionaries improves accessibility in the more specific sense for sign language users. More hypothetically, online dictionaries offer a potential accessibility advantage inasmuch as their content can be manipulated by the dictionary user. For example, when designed with accessibility in mind, webpage elements such as font sizes, line spacing, and colour schemes can be changed at the click of a button or a change in browser settings so that the information they encode is easier to perceive. Another very important potential accessibility advantage offered by online dictionaries is their potential compatibility with assistive technologies. These include technologies such as refreshable braille displays (devices that render on-screen text as braille), magnifiers (tools which make text and images on webpages appear larger), and screen readers (synthe-

sised speech tools which read the information displayed on screen). Although these possibilities exist for ensuring dictionary websites are accessible for people with visual impairments — and have been partially exploited in the case of *Mikaela Lex*, a Greek school dictionary for students with visual impairments (Gavriilidou and Garoufos 2022) — previous research suggests that the creators of many dictionary websites have not yet exploited the full potential of these accessibility features (cf. Arias-Badia and Torner 2023; Rees 2023).

This lack of accessibility is particularly troubling when the number of people with visual impairments is considered along with the social exclusion this group often suffers. Globally, approximately 285 million people have some form of visual impairment (WHO EMRO 2012). This figure is likely to increase since there is a global trend of population aging while many of the conditions that cause low vision and blindness are age-related. Generally, many people with visual impairments suffer from social exclusion as evidenced by the high rates of un- and underemployment among this cohort (cf. National Federation of the Blind 2019). Several studies have attempted to address this problem, for example through an evaluation of the accessibility of library resources (McKenzie and Casey 1998) and educational software (Dini et al. 2007). The present project continues in this vein in the hope of making a small step forward in the inclusion of people with vision impairment and blindness in professional and educational contexts.

2. Previous research on the visual accessibility of dictionary websites

To my knowledge, there are two studies that address the accessibility of dictionary websites for people with visual impairments (Arias-Badia and Torner 2023; Rees 2023). They both employ the Web Content Accessibility Guidelines (WCAG) (World Wide Web Consortium (W3C) 2023), a series of guidelines aimed at improving access for people with disabilities to websites and other web applications. The guidelines were developed as part of the Web Accessibility Initiative, which is an initiative endorsed by the W3C — the international standards organization for the World Wide Web.

The current version of the WCAG (2.2) contains thirteen guidelines organised around four key principles that describe ideal characteristics of accessible websites: Perceivable (e.g., websites should contain information such as alternative text for images), Operable (e.g., their interfaces can be accessed with both keyboard and mouse), Understandable (e.g., the language used on a website should be programmatically specified to ensure that it can be pronounced correctly by screen reading software), and Robust (e.g., content should be compatible with assistive technologies such as screen readers). Each guideline has testable *success criteria* that are graded into three conformance levels ranging from A (the minimum standard), through AA (the de facto standard), to AAA (the ideal).

The two studies mentioned differ in scope and the way in which they employ

the WCAG. Arias-Badia and Torner (2023) take a broad top-down approach, evaluating dictionary website accessibility both for people with intellectual disabilities and for people with visual impairments. Without referring to specific success criteria or conformance levels, they use the guidelines of perceivability and understandability to structure their evaluation of the influence that heading structure, use of images, and the typographic choices have on accessibility for these users. Rees (2023) adopts a narrower focus, examining the accessibility of dictionary websites for people with visual impairments. In line with common practice in web accessibility evaluations, they adopt a bottom-up approach working from dictionary websites to success criteria then to key principle. Furthermore, they evaluate the compliance of dictionary websites with success criteria related to all four key WCAG principles. The rationale is that although the Operable and Robust key principles are not directly related to visual perception, they are essential to rendering website content in a way that is perceivable and understandable for people with severe visual impairments. For example, since screen reader users often navigate by keyboard rather than mouse, websites must be operable in this way. Similarly, websites also need to be robust, inasmuch as their content should be compatible with screen readers and braille displays.

Rees (2023) employs the two-stage methodology widely used in web accessibility evaluations. Firstly, automated evaluation tools are used to scan website source code for indications of unmet success criteria. Because of the high possibility of false positives and negatives, in line with common practice (Abascal et al. 2019), a combination of three tools is used: Achecker — Web Accessibility Checker (IDI 2022), MAUVE++ (Broccia et al. 2020), and WAVE — Web Accessibility Evaluation Tool (WebAIM 2021). In a manual inspection stage, an evaluator then browses through a sample of pages from a website checking their experience against a checklist of success criteria. In their evaluation, Rees (2023) compiled a list of 46 A and AA level success criteria that were potentially applicable to dictionary sites. These were considered in both the automated evaluation and manual inspection stages. In addition, a further three potentially relevant AAA success criteria of particular significance to lexicographic applications were also considered (see supplementary information). These include "3.1.3 Unusual Words: A mechanism is available for identifying specific definitions of words or phrases used in an unusual or restricted way, including idioms and jargon"; "3.1.4 Abbreviations: A mechanism for identifying the expanded form or meaning of abbreviations is available"; and "3.1.6 Pronunciation: A mechanism is available for identifying specific pronunciation of words where meaning of the words, in context, is ambiguous without knowing the pronunciation". This two-stage methodology, the tools, and the checklist of potentially relevant success criteria are employed in the evaluations of prototype versions of EDictViz detailed in Section 3 below.

As shown in Table 1, the previous studies also differ partially in the dictionary websites they evaluate. All resources studied are monolingual. *Diccionario de la lengua española* and *Diccionario fácil* are Spanish resources while all other resources are English, although some offer localised versions² and links to

bilingual resources. All resources except *Diccionario fácil*, which is the fruit of a project aiming to serve users with cognitive impairments, are mass market resources.

Table 1: Dictionary websites included in previous accessibility evaluations

Dictionary	Publisher	URL	Arias-Badia and Torner (2023)	Rees (2023)
<i>Cambridge English Dictionary</i>	Cambridge University Press	< https://dictionary.cambridge.org/dictionary/english/ >	X	
<i>Collins Dictionary</i>	HarperCollins	< https://www.collinsdictionary.com/ >		X
<i>Diccionario de la lengua española</i>	Real Academia Española	< https://dle.rae.es/ >	X	X
<i>Diccionario fácil</i>	Plena Inclusión Madrid	< http://diccionariofacil.org/ >	X	
<i>Macmillan English Dictionary</i>	Macmillan Education	< https://www.macmillandictionary.com/ >	X	
<i>Merriam-Webster.com Dictionary</i>	Merriam-Webster	< https://www.merriamwebster.com/ >	X	X
<i>Oxford Learner's Dictionary</i>	Oxford University Press	< https://www.oxfordlearnersdictionaries.com/ >	X	

Despite their differences in scope, methodology, and the dictionaries they cover, the findings of both studies suggest that existing online dictionaries present significant accessibility challenges for users with visual impairments. These include problems with labelling, adaptable design, colour and contrast, and navigation for keyboard users, with each being discussed below.

Labelling

For users with low vision or without the sense of sight, information that is not encoded as text is imperceivable if it is not readable by a screen reader or braille display. This applies not only to the content of the page — such as the inclusion of alternative text to describe images or full-text labels for abbreviations — but also to the design of the page. Unless labelled with alternative text, a clickable arrow at the top of a page taking the user to an entry, the specific part of the dictionary page with the information about the search word, would for exam-

ple be imperceivable to many users with visual impairments. Other interactive elements such as search boxes and menus also need to be appropriately labelled. Technologically, labelling can be achieved via on-screen text, which is readable to all users, alternative text to describe images that becomes visible when a user selects or moves the mouse over an image, or Accessible Rich Internet Applications labels (ARIA-labels). These are labels in the code of a webpage that provide information about the purpose and status of interactive elements on a webpage to users of assistive technologies.

Adaptable design

One of the principal potential advantages of online dictionaries over their paper counterparts is that, in theory, they offer the user the opportunity to change the layout of the page. This does not necessarily involve including an 'Enlarge Text' or 'Change Spacing' button. Instead, pages should be coded in such a way that the user can alter the size and spacing of text programmatically (e.g., using a browser plug-in or via browser settings). To comply with the relevant WCAG success criteria at an AA level, it should be possible to enlarge text by 200 percent without distorting the rest of the page. There are similar rules for spacing. For example, it should be possible to set line height to at least 1.5 times the font size, letter spacing to at least 0.12 times the font size, word spacing to at least 0.16 times the font size, and spacing following paragraphs to at least twice the font size. In practice, making these adaptations on dictionary websites often results in a distorted unusable site.

Colour and contrast

Traditionally, colour choices in paper-based dictionaries were limited to black print on white (or slightly off-white) paper. More recently, colour printing albeit with a limited palette, became economically viable. Devices such as colour signposts have made dictionary entries easier to navigate (Dziemianko 2015; Dziemianko 2016). However, as far as people with visual impairments are concerned, the huge range of colours possible on dictionary websites presents an accessibility challenge. To comply with the WCAG at an AA level, a contrast ratio of 4.5:1 between background and text is necessary. To put this into perspective: black text on a white background has a contrast ratio of 21:1. Several of the accessibility problems observed in Rees (2023) relate to insufficient contrast in dictionary entries. Elements such as labels indicating grammatical gender, as well as illustrative examples of word use in context were especially affected by low contrast issues. There is some debate about the adequacy of the WCAG2 contrast definition, particularly its applicability to dark backgrounds (Ulitin 2023). However, for practical effects, this is not relevant here since the contrast problems observed in the previous studies of dictionary accessibility relate to light back-

grounds. More anecdotally, the figures provided in Rees (2023) suggest contrast problems that could also affect sighted users. The use of colour to encode usage warnings such as warnings about common grammatical errors (e.g., red = bad; green = good) is obviously problematic for people with colour-blindness.

Navigation for keyboard users

Many users of screen readers and other assistive technologies navigate websites using the keyboard not the mouse. Unfortunately, innovations in online dictionaries such as clickable signposts that scroll to the part of the entry of interest to the user and entries that unfold when clicked are not operable by keyboard (Rees 2023).

3. EDictViz: A potential solution

EDictViz is a website that uses a dictionary API to display dictionary data in a way that is accessible for people with visual impairments. For the purposes of prototyping, the *Merriam-Webster Advanced English Learner's Dictionary* (Merriam-Webster Dictionary API, n.d.) is employed, although any dictionary could hypothetically be used. The development of the resource follows the Agile software development methodology: an iterative process of collaborative development, evaluation and feedback, and further development in response to that feedback (Beck et al. 2001). Table 2 shows the milestones that mark each version and progress to date.

Table 2: EDictViz development plan

Version	Milestone	Achieved
0.1	Basic working prototype incorporating findings from previous research	Yes
0.2	Pass automated WCAG inspection	Yes
0.3	Pass manual WCAG inspection	Yes
0.4	Usability testing	Pending
0.5	User testing in language reception and production tasks	Pending

The next section deals with the development and evaluation using the WCAG criteria undertaken so far. Section 3.2 deals with plans for future development and evaluation. The WCAG are a good starting point for evaluating EDictViz and have the advantage of providing a broad overview of accessibility applicable

to the enormously diverse group of possible low vision users (e.g., those with low visual perception, reduced visual acuity, colour blindness, tunnel vision, etc.). However, the pending evaluation with users will be essential for establishing the effectiveness of the tool and of paramount importance to its future development.

3.1 Previous evaluation

3.1.1 Version 0.1: Visual and technical simplicity

The goal for Version 0.1 of the tool was to make a simple website that used the API to display the basic parts of a dictionary entry (headword, word class, definition, and example) when a user looked up a word. The guiding principles throughout the development process have been visual and technical simplicity. This pared-down approach has its roots in a previous evaluation of those elements on dictionary websites that cause accessibility problems for people with visual impairments. As shown in Figure 1, Rees (2023) groups these problematic elements into four categories: advertising, lexicotainment, design, and core-lexicographic. They are discussed here in descending order of the percentage of accessibility problems they represented in Rees's study.

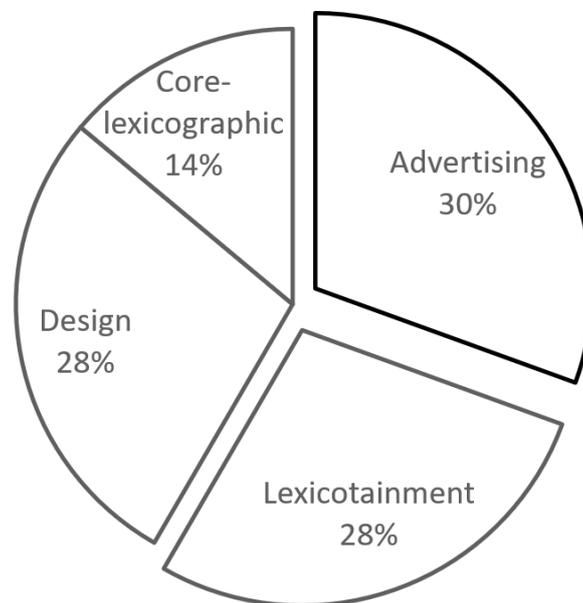


Figure 1: Sources of unmet success criteria on dictionary entry pages (data source: Rees (2023))

Advertising relates to the inclusion of advertisements and other publicity material on dictionary websites. Since advertisements are often provided by third parties, the operator of the website has limited control over their content and form. For example, advertisements are frequently missing alternative text labels, meaning that they cannot be meaningfully described by screen readers. The inclusion of advertisements often interferes with structural elements such as headings and landmarks (hidden labels for page sections), which help screen-reader users navigate the page. Although the user may not be interested in the content of the advertisements, their presence alone is disruptive to the dictionary use experience. Images, moving or still at the side or top of a screen, can be distracting to sighted users. It is reasonable to assume that this effect is multiplied if the image is blurred or otherwise partially perceivable.

More generally, empirical studies have found the presence of advertisements to be disadvantageous to efficient dictionary use. Dziemianko (2019) found a detrimental impact on performance when dictionaries were used to complete language reception, production, and vocabulary learning tasks. Advertisements on dictionary websites have also been shown to increase the time taken to resolve language doubts (Dziemianko 2019; Dziemianko 2020).

Lexicotainment — including elements such as word games, word of the day features, mailing lists, social media, and blog posts — are a common feature of dictionary websites. They often involve user interaction (e.g., clicking or dragging words with a mouse or entering an email address). They can change the content of a screen without warning (e.g., an auto-updating list of popular searches) or simply add more visually complex elements to a webpage (e.g., unlabelled images or low contrast text from social media feeds). All of this entails significant accessibility barriers.

Design relates to the structure of site and page. This includes the logical use of headings that are helpful when navigating webpages for screen-reader users. It also includes the use of labels describing regions of the page (such as the header, footer, and main section) and navigation, or text inputs such as search. Keyboard operability is a further design consideration. This includes ensuring that all links and text inputs are accessible in a logical manner using the keyboard rather than the mouse. For example, ensuring that links can be operated using the Tab and Enter keys.

Core-lexicographic issues include those affecting microstructure. These include the colour and contrast of text and background in entries. For example, markers of grammatical gender, word class, and illustrative examples are often difficult to distinguish from the page background. The use of abbreviations in entries can also pose problems for people with visual impairments. If abbreviations must be used, the full form of the abbreviated text must be provided as alternative text. It should be possible to modify the size and spacing of the text of an entry without distorting the design of the webpage.

Since EDictViz is a non-commercial project with minimal running costs,³ there is no need to include advertisements on the site. This means that 30% of the accessibility problems found in Rees (2023) can be avoided. The same logic jus-

tifies the decision to exclude lexicotainment elements, which represent a further 28% of accessibility issues found in the same study. In commercial resources, these elements are often included to decrease a website's 'bounce rate' (i.e., the percentage of users who leave a site after viewing only one page). This is a key metric for advertisers and advertising agencies. In short, the lower the bounce rate, the more attractive a site is to advertisers. This means potentially greater advertising revenues for the publisher.

At first glance, the decision to exclude advertising and lexicotainment elements seems to run counter to universal accessibility, the argument being that users with visual impairments have as much right to experience advertisements as other users. However, given the proven detrimental effect of advertisements on the speed and retention of dictionary searches, it is possible to turn this argument on its head. Truly universal access would involve removing advertisements for all users. This counterargument may also apply to lexicotainment elements, since it seems one of the principal reasons for including these elements in commercial dictionaries is to increase attractiveness of a website to advertisers.

Excluding advertising and lexicotainment allows a focus on design and core-lexicographic elements. As far as design elements are concerned, this in practice means using a colour scheme with sufficient contrast; ensuring that regions of the page such as headers, footers, navigation menus, and the main section are properly labelled; ensuring that the search box is appropriately labelled; and ensuring the site can be navigated using a keyboard only. In terms of core-lexicographic elements, this involves ensuring sufficient contrast between background and text, ensuring that all information is encoded as text rather than images, ensuring that different parts of the entry are clearly labelled, and avoiding abbreviations for metalinguistic terms. The latter two points chime with the idea of "dictionary as prose" (Hanks 1987: 118), a key idea from the user-friendly turn in lexicography of the 1980s. In practice, this means the avoidance of meta-language and the use of parenthesis, both of which are factors found to be particularly problematic in several dictionary use studies (cf. Nesi and Meara 1994; Nesi and Hail 2002).

3.1.2 Version 0.2: Passing automated WCAG evaluation

The yardstick for transition from Version 0.1 to Version 0.2 was having a site that passed automated WCAG evaluation. Passing automated WCAG inspection is understood here as having no genuine unmet success criteria indicated by the three automated evaluation tools used. Table 3 shows the unmet success criteria found for the lemma *test* with each of the three tools employed to evaluate Version 0.1 of EDictViz. The single unmet success criteria reported by Achecker is a false positive. The tool has interpreted the ARIA-label for the search form as an empty <Label> element in violation of Success Criterion 3.3.2, which is "Labels or instructions are provided when content requires user input". Issues requiring human verification refer to success criteria that cannot be evaluated by the rule-based evaluation tools (e.g., ascertaining whether alternative text accu-

rately describes an image). In this evaluation, many of the issues highlighted as requiring further evaluation are repeated false positive results related to the same lines of webpage code. This suggests an incompatibility between the way the tools search for unmet success criteria and the way EDictViz is programmed. In any case, issues requiring human verification are covered by the manual inspection stage of evaluation discussed in Section 3.1.3 below.

Table 3: Results of automated evaluation

Evaluation tool	Unmet success criteria	Requiring human verification
Achecker	1	50
Mauve++	0	27
WAVE	0	0

Figure 2 from WAVE provides a visual summary of the accessibility of this version. WAVE detects no unmet success criteria ('Errors' and 'Contrast Errors'). There are no potential issues requiring human evaluation ('Alerts'). The label specifying the language of the page is noted as an accessibility feature. The 24 Structural Elements on the page are highlighted. These include the labelled header and footer sections, search box, navigation menus, main section, and hierarchically ordered headings, all of which are helpful to users navigating with screen readers. This is also true of the two ARIA-labels highlighted.

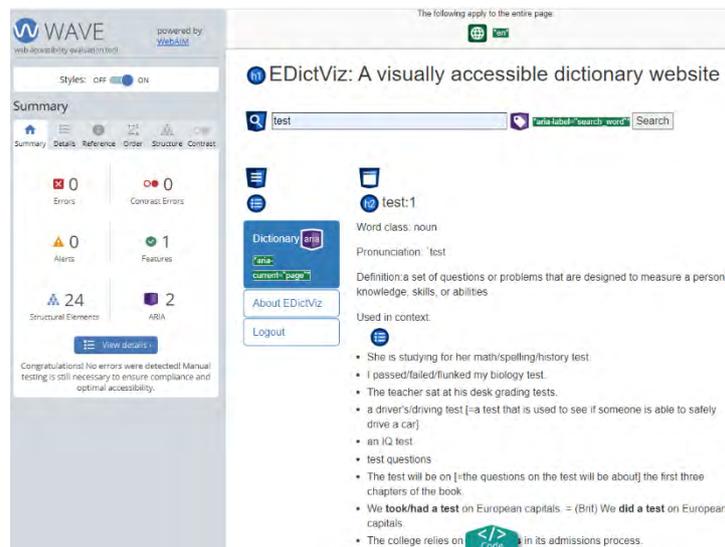


Figure 2: Visual summary of the accessibility of EDictViz Version 0.2

3.1.3 Version 0.3: Passing manual inspection

The results of the manual inspection of Version 0.2 of EDictViz suggest the pared-down design philosophy offers accessibility advantages. Only 67% (33 of 49) of dictionary webpage success criteria applied to this version of EDictViz. Of these 33 applicable success criteria, only two are clearly unmet by EDictViz. These include Success Criterion 2.4.1, namely "Bypass Blocks: Provide a 'Skip to Content' link", and Success Criterion 2.4.5, which is "Multiple Ways: Offer several ways to find pages" (see supplementary information).

The 'Skip to Content' issue is remedied in Version 0.3 where a link is provided at the top of each page to allow the user to skip directly to the dictionary entry. The idea of allowing the user to skip to the content of interest to them has parallels with what was once considered a user-friendly innovation in learners' dictionaries (cf. Gouws 2014). Some learners' dictionaries provide maps or tables of contents for longer entries. In paper-based resources, these are numbered lists of the senses of a word included in the entry. The user can quickly skip to the relevant sense without having to read through the whole entry. In several commercial online dictionaries these entry maps are hyperlinked to the relevant sense. Unfortunately, this feature is often implemented in such a way that it cannot be perceived or operated by users with visual impairments (Rees 2023).

In Version 0.2 the only way to find a word in EDictViz is to search for it in the search box. This runs contrary to Success Criterion 2.4.5 "Multiple Ways: Offer several ways to find pages". To remedy this, an alphabetic index of headwords is included in Version 0.3. Here again there are parallels between this accessibility feature and lexicographic practice. Freedom from the constraints of the alphabetical index has been a major hope for many lexicography researchers in the digital age (De Schryver 2003). However, this somewhat maligned feature still has its uses. Due to limitations in the methods used to compile digital dictionaries (cf. Frankenberg-Garcia et al. 2021), users searching for derived or compound forms of a word are often left unsatisfied. For example, searching for the adjective *assumed* in the *Cambridge English Dictionary* takes the user to the page for the verb *assume*. However, the alphabetical index in the 'Browse' box at the bottom of the page contains the adjectival uses *assumed debt*, *assumed liabilities*, and *assumed name*, which may help the user resolve their query (Rees 2024).

3.2 Future development and evaluation

Using the WCAG to evaluate the accessibility of dictionary websites for people with visual impairments has several advantages. There are several types of visual impairments, many of which bring different accessibility needs. The WCAG approach allows an abstract evaluation over several visual impairments, concentrating on accessibility needs that are common to many of them. In this way, the approach addresses the needs of a greater range of users than would be feasible with research studies that directly involved users. Moreover, it is a widely employed methodology that is relatively cheap and straightforward to

put into practice. Furthermore, it arguably offers an ecological validity advantage over tests carried out in closely controlled conditions.

Despite these clear advantages, the WCAG are not the perfect solution. As far as is practicable, there is an obvious need for end users with low vision to be directly involved in the development and evaluation of tools which aim to serve them. This may allow the investigation of other qualitative aspects of user experience not addressed by the WCAG evaluation. For example, it may provide further insight into user perceptions of the cognitive load involved in using the tool, the intuitiveness of its user interface, or other subjective measures of usability, such as the emotional impact of the tool. Future development and evaluation of EDictViz will therefore involve people with visual impairments more directly.

Two key aspects of the tool need further evaluation. Firstly, its usability for people with visual impairments needs to be investigated in comparison with other online dictionaries. Secondly, its effectiveness in typical tasks for which dictionaries are employed needs to be evaluated in comparison with other online dictionaries. Plans for studies with users to respond to these needs are set out in Sections 3.2.1 and 3.2.2, respectively.

A key challenge in carrying out dictionary use research is selecting vocabulary that users do not already know but could reasonably be expected to need in real life. Academic vocabulary lists provide a solution (Coxhead 2000; Rees 2021). Since many users of monolingual learners' dictionaries are university students or staff who use English for academic purposes, it is reasonable to assume that they will need to use the words on these lists. Furthermore, the words on these lists can be ordered by the frequency at which they are found in language corpora — large collections of machine-readable text. Following usage-based theories of language learning, learners are less likely to know words that appear at lower frequencies.

Recruiting low vision users of academic English represents a logistical challenge. Firstly, they are a relatively small, though nonetheless significant, subset of the population. Secondly, low vision users are a heterogeneous group (e.g. including those with low visual perception, those with reduced visual acuity, colour blindness, or tunnel vision). A key design assumption of EDictViz is that there is a limit to the specificity with which these user groups can be targeted. In practical terms, this means that it is not considered feasible to build individual dictionaries for those who need to adjust contrast, text size and spacing, those who require magnifiers, and those who require screen readers. EDictViz attempts to cater to all these groups. It is, however, considered feasible to recruit 10 to 12 users per study who represent this range of use cases for each of the studies outlined below.

3.2.1 Version 0.4: Usability testing

In this experimental study, low vision users will look up the meaning of two comparable sets of lower-frequency words from the New Academic Wordlist (Cox-

head 2000). After doing this, they will complete the System Usability Scale (SUS) questionnaire (Brooke 1996). The SUS is used to calculate a score representing the usability of the tool. The decision to use the SUS, a proxy for usability, rather than a measure of accessibility is firstly based on the premise that usability and accessibility are mutual requirements (e.g., a tool cannot be usable if it is not accessible, nor can it be accessible if it is not usable). It is also based on the widespread use, ease of administration in challenging environments, and scalability of the SUS. The SUS is a de facto standard measure in usability testing. It is referenced in over 1,300 publications, including those related to the evaluation of lexicographic tools (Frankenberg-Garcia et al. 2019). It has proven quick, cheap to administer and, importantly in this context, reliable even with the expected small sample sizes given the difficulties associated with recruiting low vision users and the diversity of this group's needs. The SUS will allow the collection of subjective measures of the tool's usability that it was not possible to capture with previous WCAG-based evaluations.

In line with common practice, follow-up interviews will complement the SUS questionnaire and allow further insight into user experiences of EDictViz and opportunities for its improvement (e.g. how to make the interface more intuitive and enjoyable to use). It is important to clarify that the study has not been completed yet and that even if it had been, it would be beyond the scope of this article to report it in sufficient detail.

3.2.2 Version 0.5: Vocabulary knowledge testing

This experimental study is designed to compare low-vision users' productive and receptive vocabulary knowledge changes when using EDictViz with those made using a traditional online monolingual English learners' dictionary. This will go some way to elucidating the extent to which EDictViz mitigates barriers to successful dictionary use for people with visual impairments.

The study will use a counterbalanced design in which participants complete two tasks in two conditions: (1) using a standard online dictionary, and (2) using EDictViz. The counterbalanced design has ethical advantages since it ensures that no group is disadvantaged by missing out on any potential benefits of each condition. The first task is intended to measure receptive vocabulary gain from reading. In this task, participants look up the meaning of two comparable sets of lower-frequency words from the New Academic Wordlist (Coxhead 2000) in the two conditions. The second task is intended to measure vocabulary gain for writing. Participants complete two comparable sets of sentence frames (sentences with a missing word) with salient academic collocations — words that it has become conventional to use together (e.g., *powerful computer* but not *strong computer*) (Roberts et al. 2017). In both tasks participants self-rate their knowledge of the vocabulary items using a measure based on the Test of Academic Lexicon (TAL) (Scarcella and Zimmerman 1998). Once the tasks are completed, scores are calculated using acceptable responses to the sentence frames and TAL level

in each condition. Scores are then compared for each condition.

Like the usability study outlined above, insight into user experiences of EDictViz and opportunities for improvement will be garnered from follow up interviews. The proposed study is in the design stage. If results were available, it would be beyond the scope of this article to provide the required context and discussion.

4. Discussion and conclusions

In this account of the development of EDictViz, it has been argued that dictionary access is important not only because of the functional role that dictionaries play as information tools, but also because of their potential as motors of social inclusion for marginalised groups such as persons with visual impairments. Unfortunately, paper-based dictionaries have historically proved difficult to access for people with visual impairments. Furthermore, it has been shown that while online dictionary websites, in theory, offer greater accessibility, they present significant accessibility challenges in practice, as demonstrated by previous research. EDictViz was posited as an attempt to remedy this. The project so far has not only developed an ostensibly more accessible online dictionary but also highlighted some of the communalities between access as understood in lexicography and accessibility studies. It could act as a model for the application of many of the guidelines from previous research on dictionary website accessibility. This has the potential to improve both access to dictionary websites and other online lexicographic resources for persons with visual impairments, as well as for people in general.

To understand the wider implications of the project, it is helpful to keep in mind Greco's (2018: 211) three shifts that mark the entrance and development of the construct of accessibility in various research fields. One key shift is from a reactive approach where access is 'bolted on' as an afterthought to existing products or services, to a proactive one where products and services are designed with access in mind from the start. Another is the shift from a maker-centred approach to access (in which the makers of a product or service make assumptions about users' accessibility needs) to one where users are active stakeholders in the design process. A further shift is the one from a particularist account of accessibility, in which access concerns concrete groups of people who were often persons with disabilities, to one where access concerns all users.

EDictViz clearly represents a proactive approach to access. It is designed from the ground up with accessibility in mind. Although many existing dictionary websites have accessibility features indicating that their designers are aware of the need to ensure accessibility, the effectiveness of these features is counteracted by overall inaccessible design.

Regarding the next shift, the position of EDictViz is less clear. Thus far, the development of EDictViz has exemplified a maker's account of access. Methodologically, the reliance on the WCAGs rather than direct testing with users entails possibly unfounded assumptions about the needs of persons with visual im-

pairments. Moreover, as discussed above, it does not provide insight into user perceptions of the tool, such as how enjoyable or intuitive it is to use or even how useful it is. The planned development and evaluation of future versions of the tool will reveal whether the assumptions about the needs of persons with visual impairments are justified. It will also mark a user-centred shift in the tool's development. The SUS study will provide a subjective measure of usability, and the vocabulary knowledge study will provide a measure of the tool's effectiveness. The follow up interviews in both studies will allow the intended audience to give their own account of their lexicographic needs and to indicate the extent to which the tool meets them. More generally, it is hoped that these future studies will elucidate the extent to which, if at all, designing lexicographic resources for accessibility implies compromising the richness of lexicographic content. They may for example shed light on how or if visual indicators of word frequency can be made perceptible to screen reader users.

Although EDictViz embodies a particularist account of access, aimed as it is at persons with visual impairments, it also points the way to a universalist one. This is achieved not only by providing a model for making dictionaries more visually accessible to people in general, but also by implementing good practices from research about the most efficient ways to access lexicographic information. Thus far the project, with its WCAG-based evaluation, has only scratched the surface of the opportunities for mutually beneficial exchange between the fields of accessibility studies and lexicography. It is hoped that further development of the project, including the proposed studies with persons with visual impairments outlined above, will spark yet more progress towards the universal accessibility of lexicographic information.

Acknowledgments

I gratefully acknowledge the support of the BALEAP ('The global forum for EAP professionals') 2021 Funding Stream ('Social Justice in EAP') without which this project would not have been possible. I would also like to thank the two adjudicators for their careful reading and encouraging comments.

Endnotes

1. Development Version 0.3 of the tool is found at <https://edictviz.com/>. Access is available on request.
2. These versions contain elements such as introductory information and grammatical labels in the language of the user's region if that differs from the language of the dictionary. For example, a user from France using *Collins English Dictionary* would initially be presented with this information in French.
3. The first development version of EDictViz went live in June 2022. The initial development costs and hosting costs for a period of four years were financed by a small research grant from the BALEAP Funding Stream 2021 'Social Justice in EAP'.

References

- Abascal, J., M. Arrue and X. Valencia. 2019. Tools for Web Accessibility Evaluation. Yesilada, Y. and S. Harper (Eds.). 2019. *Web Accessibility: A Foundation for Research*: 479-503. London: Springer.
- Arias-Badia, B. and S. Torner. 2023. Bridging the Gap between Website Accessibility and Lexicography: Information Access in Online Dictionaries. *Universal Access in the Information Society* 23(2): 545-560.
- Beck, K., M. Beedle, A. van Bennekum, A. Cockburn, W. Cunningham, M. Fowler, J. Grenning, J. Highsmith, A. Hunt, R. Jeffries, J. Kern, B. Marick, R.C. Martin, S. Mellor, K. Schwaber, J. Sutherland and D. Thomas. 2001. *Manifesto for Agile Software Development*. <http://agilemanifesto.org/> [21 June 2020]
- Broccia, G., M. Manca, F. Paternò and F. Pulina. 2020. Flexible Automatic Support for Web Accessibility Validation. Lampe, C., J. Nichols, K. Karahalios, G. Fitzpatrick, U. Lee, A. Monroy-Hernandez and W. Stuerzlinger (Eds.). 2020. *Proceedings of the ACM on Human-Computer Interaction* 4(EICS): 1-24.
- Brooke, J. 1996. SUS: A Quick and Dirty Usability Scale. Jordan, P.W., B. Thomas, I.L. McClelland and B. Weerdmeester (Eds.). 1996. *Usability Evaluation in Industry*: 189-194. London: Taylor & Francis.
- Coxhead, A. 2000. A New Academic Word list. *TESOL Quarterly* 34(2): 213-238.
- De Schryver, G.-M. 2003. Lexicographers' Dreams in the Electronic-Dictionary Age. *International Journal of Lexicography* 16(2): 143-199.
- Dini, S., L. Ferlino, A. Gettani, C. Martinoli and M. Ott. 2007. Educational Software and Low Vision Students: Evaluating Accessibility Factors. *Universal Access in the Information Society* 6(1): 15-29.
- Dziemianko, A. 2015. Colours in Online Dictionaries: A Case of Functional Labels. *International Journal of Lexicography* 28(1): 27-61.
- Dziemianko, A. 2016. An Insight into the Visual Presentation of Signposts in English Learners' Dictionaries Online. *International Journal of Lexicography* 29(4): 490-524.
- Dziemianko, A. 2019. The Role of Online Dictionary Advertisements in Language Reception, Production, and Retention. *ReCALL* 31(1): 5-22.
- Dziemianko, A. 2020. Smart Advertising and Online Dictionary Usefulness. *International Journal of Lexicography* 33(4): 377-403.
- Dziemianko, A. 2022. The Usefulness of Graphic Illustrations in Online Dictionaries. *ReCALL* 34(2): 218-234.
- Frankenberg-Garcia, A., G.P. Rees and R. Lew. 2021. Slipping Through the Cracks in e-Lexicography. *International Journal of Lexicography* 34(2): 206-234.
- Frankenberg-Garcia, A., G.P. Rees, R. Lew, J. Roberts, N. Sharma and P. Butcher. 2019. ColloCaid: A Tool to Help Academic English Writers Find the Words They Need. Meunier, F., J. van de Vyver, L. Bradley and S. Thouësy (Eds.). 2019 *CALL and Complexity — Short Papers from EUROCALL 2019*: 144-150. s.l.: Research-publishing.net.
- Gavriilidou, Z. and A. Garoufos. 2022. The Lexicographic Protocol for Mikaela-Lex: A Free Online Dictionary of Greek Accessible for Visually-Impaired Senior Elementary Children. Klosa-Kückelhaus, A., S. Engelberg, C. Möhrs and P. Storjohann (Eds.). 2022. *Dictionaries and Society. Proceedings of the XX EURALEX International Congress, 12–16 July 2022, Mannheim, Germany*: 563-568. Mannheim: IDS-Verlag.

- Gavriilidou, Z., M. Mitsiaki and A. Fliatouras (Eds.).** 2021. *Proceedings of the XIX EURALEX International Congress, 7–9 September 2021, Virtual. Volume 1: Lexicography for Inclusion*. Komotini, Greece: European Association for Lexicography.
- Gouws, R.H.** 2014. Article Structures: Moving from Printed to e-Dictionaries. *Lexikos* 24: 155-177.
- Greco, G.M.** 2018. The Nature of Accessibility Studies. *Journal of Audiovisual Translation* 1(1): 205-232.
- Hanks, P.** 1987. Definitions and Explanations. Sinclair, J. (Ed.) 1987. *Looking Up: An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary*: 115-136. London/Glasgow: Collins ELT.
- IDI.** 2022. Web Accessibility Checker for WCAG 2.0 AODA Compliance Monitoring. Toronto: Inclusive Design Institute.
<https://achecker.achecks.ca/checker/index.php> [15 July 2022]
- Jackson, H.** 2018. English Lexicography in the Internet Era. Fuertes-Olivera, P. (Ed.). 2018. *The Routledge Handbook of Lexicography*: 540-553. London: Routledge.
- Lew, R.** 2011. Space Restrictions in Paper and Electronic Dictionaries and Their Implications for the Design of Production Dictionaries. Bański, P. and B. Wójtowicz (Eds.). 2011. *Issues in Modern Lexicography*. München: Lincom Europa.
- Lew, R.** 2012. How Can We Make Electronic Dictionaries More Effective? Granger, S. and M. Paquot (Eds.). 2012. *Electronic Lexicography*: 343-362. Oxford: Oxford University Press.
- Lew, R.** 2015a. Dictionaries and Their Users. Hanks, P. and G.-M. de Schryver (Eds.). 2015. *International Handbook of Modern Lexis and Lexicography*: 1-9. Berlin/Heidelberg: Springer.
- Lew, R.** 2015b. Research into the Use of Online Dictionaries. *International Journal of Lexicography* 28(2): 232-253.
- McKee, R.L. and D. McKee.** 2013. Making an Online Dictionary of New Zealand Sign Language. *Lexikos* 23: 500-531.
- McKenzie, E.M. and K.E. Casey.** 1998. Using Adaptive Technology to Provide Access to Blind, Low-Vision, and Dyslexic Patrons. *Law Library Journal* 90(2): 157-182.
- Merriam-Webster Dictionary API.** n.d.
<https://dictionaryapi.com> [12 July 2023]
- National Federation of the Blind.** 2019. Blindness Statistics. *National Federation of the Blind*.
<https://nfb.org/resources/blindness-statistics> [25 March 2022]
- Nesi, H.** 1998. Defining a Shoehorn: The Success of Learners' Dictionary Entries for Concrete Nouns. Atkins, B.T.S. (Ed.). 1998. *Using Dictionaries: Studies of Dictionary Use by Language Learners and Translators*: 159-178. Tübingen: Niemeyer.
- Nesi, H. and R. Haill.** 2002. A Study of Dictionary Use by International Students at a British University. *International Journal of Lexicography* 15(4): 277-305.
- Nesi, H. and P. Meara.** 1994. Patterns of Misinterpretation in the Productive Use. *System* 22(1): 1-15.
- Rees, G.P.** 2021. Discipline-Specific Academic Phraseology: Corpus Evidence and Potential Applications. Charles, M. and A. Frankenberg-Garcia (Eds.). 2021. *Corpora in ESP/EAP Writing Instruction: Preparation, Exploitation, Analysis*: 32-54. London: Routledge.
- Rees, G.P.** 2023. Online Dictionaries and Accessibility for People with Visual Impairments. *International Journal of Lexicography* 36(2): 107-132.
- Rees, G.P.** 2024. Academic Word Families in Online English Dictionaries. *Lexikos* 34: 437-468.
- Rees, G.P.** 2025. Multimedia in Dictionaries. Nesi, H. and P. Milin (Eds.). 2025 *International Encyclopedia of Language and Linguistics*. Third edition. (Reference Module in Social Sciences). Amsterdam: Elsevier.

- Rees, G., B. Arias-Badia, E. Bernal and S. Torner.** 2024. Lexicography and Accessibility. Despot, Kristina Š., Ana Ostroški Anić and Ivana Brač (Eds.). 2024. *Pre-conference Workshop at the XXI EURALEX International Congress, Lexicography and Semantics, Cavtat, Croatia, 8–12 October, 2024.*
- Roberts, J.C., A. Frankenberg-Garcia, R. Lew, G.P. Rees and J. Pereda.** 2017. Visualisation and Graphical Techniques to Help Writers Write More Idiomatically. 2017 Poster presented at IEEE VIS 2017. Phoenix, Arizona: IEEE.
- Scarcella, R. and C. Zimmerman.** 1998. Academic Words and Gender: ESL Student Performance on a Test of Academic Lexicon. *Studies in Second Language Acquisition* 20(1): 27-49.
- Ulitin, K.** 2023. Applying APCA and Huetone for Color Accessibility of User Interfaces. Stephanidis, Constantine, Margherita Antona, Stavroula Ntoa and Gavriel Salvendy (Eds.). 2023. *HCI International 2023 Posters*: 382-388. Cham: Springer Nature.
- WebAIM.** 2021. WAVE Evaluation Tool. Logan, UT: WebAIM.
<https://wave.webaim.org/>
- WHO EMRO.** 2012. WHO Releases New Global Estimates on Visual Impairment. *World Health Organization — Regional Office for the Eastern Mediterranean.*
<http://www.emro.who.int/control-and-preventions-of-blindness-and-deafness/announcements/global-estimates-on-visual-impairment.html> [13 July 2023]
- World Wide Web Consortium (W3C).** 2023. Web Content Accessibility Guidelines (WCAG) 2.2.
<https://www.w3.org/TR/WCAG22/> [6 August 2025]

Supplementary Information

The supplementary information can be found at:
https://osf.io/75zum/?view_only=d6d20a8164fe4af0b4a96c3607871870

One Name, Two Genres: The Curious Case of Polish Encyclopaedias of Law

Agnieszka Rzepkowska, *Faculty of Humanities, Institute of Linguistics and Literary Studies, University of Siedlce, Siedlce, Poland*
(agnieszka.rzepkowska@uws.edu.pl)
(<https://orcid.org/0000-0003-3381-0870>)

Abstract: This study examines Polish legal reference works through systematic analysis of publications labelled as encyclopaedias of law, addressing a gap in legal metalexigraphy. It sheds some light on the typology and characteristics of such volumes. The library survey revealed two types of Polish encyclopaedias of law: lexicographic works with alphabetically ordered entries presenting fragmented knowledge and monographic works structured in chapters that resemble academic textbooks. In fact, in the majority of cases, only opening a given work lets the user learn what type of publication it is. The study describes 78 works published between 1842 and 2023 in terms of the publication year, the fields of law covered, size, scope and target users. The analysis demonstrates that lexicographic encyclopaedias of law predominantly specialise in specific legal fields and employ various systematisation techniques including cross-referencing and thematic indices, while monographic encyclopaedias of law primarily address general Polish law for educational purposes. An important observation emerging from this study is that monographic publications, despite their textbook-like structure, share fundamental characteristics with lexicographic ones and serve complementary knowledge-organizing functions that demonstrate lexicographic adaptability to diverse user needs within specialized domains. The findings challenge traditional lexicographic classification systems by suggesting that functional criteria may be more significant than formal structural features in understanding specialised lexicographic practice.

Keywords: ENCYCLOPAEDIA OF LAW, LEXICOGRAPHY, POLISH LEXICOGRAPHY, LEGAL LANGUAGE, LSP, SPECIAL-PURPOSE DICTIONARY

Opsomming: Een naam, twee genres: Die unieke geval van Poolse regsensiklopedieë. Hierdie artikel ondersoek Poolse regsensiklopedieë deur 'n sistematiese analise van publikasies wat as regsensiklopedieë geëtiketteer is, uit te voer, en wys 'n gaping in die regsmetaleksikografie uit. Dit belig die tipologie en eienskappe van hierdie bronne. Die biblioteekopname het twee tipes Poolse regsensiklopedieë geïdentifiseer: leksikografiese werke met alfabeties geordende inskrywings wat gefragmenteerde kennis aanbied en monografiese werke wat in hoofstukke verdeel is wat soos akademiese handboeke lyk. In die meeste gevalle besef die gebruiker in werklikheid eers watter tipe publikasie dit is wanneer die gegewe werk oopgeslaan word. Die studie beskryf 78 werke wat tussen 1842 en 2023 gepubliseer is volgens die publikasiejaar, regsvelde wat gedek word, grootte, omvang en teikengebruikers. Die analise toon dat leksikografiese regsensiklo-

pedieë hoofsaaklik in spesifieke regsvelde spesialiseer en verskeie sistematiseringsstegnieke, insluitend kruisverwysing en tematiese indeksering, aanwend, terwyl monografiese regsensiklopedieë hoofsaaklik die algemene Poolse reg vir opvoedkundige doeleindes aanspreek. 'n Belangrike waarneming wat uit hierdie studie voortspruit, is dat monografiese publikasies, ongeag hul handboekagtige struktuur, fundamentele eienskappe met leksikografiese publikasies deel en aanvullende kennisorganiserende funksies verrig wat leksikografiese aanpasbaarheid by diverse gebruikersbehoefes binne gespesialiseerde velde vertoon. Die bevindings daag tradisionele leksikografiese klassifikasiesistels uit deur te suggereer dat funksionele kriteria dalk belangriker as formele strukturele kenmerke in die begrip van die gespesialiseerde leksikografiese praktyk kan wees.

Sleutelwoorde: REGSENSIKLOPEDIE, LEKSIKOGRAFIE, POOLSE LEKSIKOGRAFIE, REGSTAAL, TSD, SPESIALE DOEL WOORDEBOEK

1. Introduction

Legal language has attracted considerable attention from lexicographers and terminographers, resulting in a wide array of lexicographic works, from glossaries and dictionaries to lexicons and encyclopaedias. While the various names these works are given suggest genre differences, in practice it is often impossible to predict the content, functions and structure of such works without consulting them directly. This study looks at publications labelled as encyclopaedia of law in order to identify their genre characteristics.

This study arose from the observation that, in Poland, publications labelled as encyclopaedias of law frequently diverge in nature. Some volumes follow traditional lexicographic models with alphabetically organised entries presenting fragmented knowledge (hereinafter "lexicographic encyclopaedias of law"). Others resemble monographs with chapter-based structures offering systematic surveys of legal topics (hereinafter "monographic encyclopaedias of law"). Many of these monographic works function as academic textbooks for non-law students and are aligned with an academic course traditionally known as Encyclopaedia of Law¹. This duality raises a scholarly question: do these monographic works, despite not being conceived as lexicographic encyclopaedias, share features with thematic legal encyclopaedias? Unlike legal encyclopaedic traditions in most other countries where terminological distinctions clearly separate reference works from academic textbooks, the Polish practice of using identical nomenclature for functionally distinct publications appears to be distinctive.

This paper reviews Polish library holdings for works titled "encyclopaedia of law" (hereinafter "encyclopaedias of law"), analyses their typological features and examines how monographic volumes align with encyclopaedic and lexicographic concepts. In doing so, the study seeks to contribute to the relatively underexplored area of descriptive monolingual legal metalexigraphy in Poland, offering insights into the lexicographic and monographic forms of encyclopaedias of law.

Research on legal lexicography and terminography is extensive but predominantly focuses on legal dictionaries rather than encyclopaedias. While bibliographies of law dictionaries and encyclopaedias exist (e.g., De Groot and Van Laer's bibliography of bilingual and multilingual legal dictionaries (2005)), they often form part of larger general lexicographic projects (e.g., Wojan's bibliography of translation dictionaries and popular encyclopaedias (2013)). Beyond bibliographic work, research in this field has focused on areas such as term presentation (Bajčić 2017), the quality of legal dictionaries (De Groot and Van Laer 2006; 2008), challenges in compiling legal dictionaries (Greenberg 2014), the lexicographic presentation of legal collocations (Micha and Mroczynska 2022) and translation applicability (Šarčević 1989; Chromá 2004; Szemińska 2011, 2014; Nielsen 2010). However, monolingual legal terminography has received limited attention, with researchers focusing primarily on bilingual and multilingual studies (e.g., Šarčević 1988; Nielsen 1994; De Groot and Van Laer 2005). One exception is a study into English and Russian monolingual terminological dictionaries by Bezuglova and Kuznetsova (2016). This review of encyclopaedias of law aims to fill this gap by providing insights into the Polish contemporary and past publishing practices and the distinctive conceptualisation of encyclopaedias of law in Poland.

2. Encyclopaedia as a lexicographic work of reference

Encyclopaedias are among the works of reference that lexicography deals with. Until the end of the 18th century, they were regarded as varieties of dictionaries. Now, encyclopaedias are sometimes classified next to dictionaries and all other works of reference, or alternatively as types of dictionaries, with entries typically in alphabetical order (McArthur 1986: 103). Contemporary works often blur the boundaries, with some encyclopaedic works being called dictionaries, such as *The Oxford Dictionary of Natural History* (Allaby 1985) (McArthur 1996: 323, Bergenholtz 2012: 22). Encyclopaedias range from single volumes (e.g., *The Cambridge Encyclopaedia of Language* (Crystal 2010)) to multi-volume undertakings (e.g., 32 volumes of *Encyclopaedia Britannica*) (McArthur 1996: 323).

The scholarly debate around distinguishing encyclopaedias from dictionaries centres on several key criteria, particularly the type of knowledge presented: linguistic in dictionaries and encyclopaedic in encyclopaedias (Bergenholtz and Tarp 1995, Bajčić 2017: 55-56). Geeraerts suggests that the distinction lies in macrostructural and microstructural differences:

Macrostructurally, the encyclopedia focuses on proper names, nouns, and maybe a number of other elements from open word classes, whereas the dictionary includes all word classes (typically excluding all or most proper names). Microstructurally, the encyclopedia focuses on expert information as provided by scientific, technical, or professional experts, whereas that information is only one of the types of semantic description that the dictionary may include, together with the more everyday uses of the words. (Geeraerts 2007: 1164)

However, this distinction remains contentious. Allan treats encyclopaedias as structured databases containing comprehensive information across multiple areas of human activity, with dictionaries serving as access points to encyclopaedic entries (Allan 2006: 573). Haiman controversially argues that dictionaries are encyclopaedias, seeing only the frequency of proper names as distinguishing them (Haiman 1980: 331). The subsequent debate between Haiman (1982) and Frawley (1981) identifies typical encyclopaedic traits: emphasis on cultural rather than linguistic knowledge, objective rather than subjective facts, pragmatics over semantics, synthetic over analytic knowledge and proper names over common names (Haiman 1980: 331-354; Frawley 1981: 53).

Hartmann and James define an encyclopaedia as

a type of reference work which presents factual information in a wide range of subject disciplines. This information is collected from expert consultants and arranged in a combination of systematic and alphabetical order for the benefit of the general educated reader. Encyclopedic definitions are more detailed and less vague than the meaning explanations offered in general dictionaries (Hartmann and James 2002: 48).

They emphasise that encyclopaedic definitions reflect encyclopaedic rather than linguistic knowledge, though these categories often overlap in practice. Encyclopaedic information "is more concerned with the description of objective realities than the words or phrases that refer to them" (Hartmann and James 2002: 49).

Both dictionaries and encyclopaedias share systematic structures enabling quick information retrieval, typically through alphabetical arrangement — a method nearly 1,000 years old for encyclopaedias. Before printing came to Europe, encyclopaedias had been arranged systematically by subject. In Roman times the approach taken was practical, depending on what was seen as the most urgent everyday topic, this is what the encyclopaedia started with. Some modern works maintain this ancient arrangement (Collison and Preece 2024), such as the latest edition of *Encyclopaedia Britannica* (32 volumes), which combines three approaches: two alphabetical (*macroaedia* and *microaedia*) and one thematic (*propaedia*) (Peters 2015: 73-74).

Lexicographic works are to serve two main functions: communication-oriented and cognition-oriented. Dictionaries primarily address text understanding and production in specific communicative situations, while encyclopaedias and encyclopaedic dictionaries provide knowledge without communicative context to help solve problems (Bergenholtz and Nielsen 2006: 286; Ptaszynski 2010: 412). From a lexicographical viewpoint, understanding a work's function and purpose is essential for selecting appropriate contents and presentation methods, since user utility is paramount (Louw 1999: 286; Tarp 2000: 204; Nielsen 2008: 187-188). This functional understanding necessitates specialised dictionaries and encyclopaedias for particular knowledge fields. As Allan puts it, "[m]ultiple dictionaries and encyclopaedias would model individual human capacities and divide data and processing into manageable chunks" (Allan 2006: 574).

2.1 Encyclopaedia of law

Elaborating on the classification of specialised lexicographic works, Bergenholtz and Tarp suggest their division into dictionaries (with information about words), encyclopaedias (with information about facts) and encyclopaedic dictionaries (with information about words and facts). However, they emphasise that the information included depends on field, purpose and target group, ultimately affecting dictionary structure (Bergenholtz and Tarp 1995: 28-30). Between dictionaries and encyclopaedias lies the category of lexicons, which combines features of both reference types (Hartmann and James 2002: 86, Lukszyn 1998: 286, 2005: 63). Lexicons typically present alphabetically arranged entries with both linguistic and encyclopaedic information, making them particularly relevant to specialised domains where terminological precision and factual context are equally important. This intermediate category is especially pertinent to legal reference works, where the boundary between linguistic and encyclopaedic knowledge often becomes blurred, as evidenced in the Polish publications examined in this study.

Encyclopaedias of law are examples of lexicographic works dealing with a specific area of human activity which is highly culturally bound (Bergenholtz and Nielsen 2002: 15).

[A]n encyclopaedia of law with pretensions to completeness is a huge and complex undertaking. In its encyclopaedic mode, the 'legal' could include both a doctrinal analysis of the questions behind every legal term and an account of every object, process or phenomenon to which law at one time or other has directed its gaze. (Hutton 2014: 180)

Legal lexicography typically follows the Roman *definitions*, a definition type that is not purely lexical, as it never aims at presenting the terminological meaning of terms only, but attempts to adopt a descriptive and synthetic approach (Barenot 2014: 14). Similarly to what has been said about encyclopaedias and dictionaries in general, Hayaert notices that in the legal context there is no "waterproof division" (Hayaert 2011: 316, as cited in Barenot 2014: 14) between a dictionary of language and encyclopaedia either. There are even fewer differences between an LSP² dictionary and an encyclopaedia, as the former rarely focuses on linguistic aspects of headwords but more on technical, professional and scientific expert knowledge. Both lexicographic works may be large-scale publications with thousands of entries or small-scale with an emphasis on the most frequent and most central senses (Geeraerts 2007: 1164-1165), both may cover a legal system as a whole, a narrow area of law or a single legal theme, and both may be alphabetical, systematic or a bit of each (Mattila 2016: 37; Nielsen 2023: 434). The traditional metalexicographic divisions have become less clear owing to the requirements of the legal matter and the practicality expected from lexicographic works. But as an encyclopaedia tends to be seen as a means of serving "broader cultural needs" (Louw 1999: 285), no wonder that it has found a place in describing legal issues.

3. Encyclopaedia of law as a monographic publication

Monographic encyclopaedias of law seem to have much in common with ancient encyclopaedias arranged by subject and importance. In the introduction to his monographic encyclopaedia of law, Ahrens (1862: 1) references the definition of encyclopaedia as a systematic review of a field or fields of knowledge in its entirety. Other scholars agree with that when explaining why they call their academic books encyclopaedias of law: first of all, Encyclopaedia of Law³ is the name of an academic course encompassing a few fields of law taught at non-legal studies and, secondly, the information presented at such courses is encyclopaedic in nature, that is, it often skips details and focuses on presenting a general image of a given institution of law (e.g. *Encyklopedia prawa ...* 1917: 7; Łopatka 1995a: 7; Siuda 2010: 11). Therefore, Encyclopaedia of Law, as the name of an academic course, is a review of individual fields of law which, when put next to each other, depict the essence of what is most important in law. Additionally, such texts should concentrate on basic concepts and define them in the simplest possible way. At the same time any comments and clarifications need to be limited to minimum, which will let the lecturer decide on the number of details included in the course (Izdebski and Murzyn 1968: 11). Encyclopaedia of Law is also seen as an introductory course preparing students for other, more detailed lectures on law, such as civil law or commercial law (Lewandowski 2003: 7).

Academic books for Encyclopaedia of Law courses have evolved during their presence at Polish universities. Initially, they were for part-time students and were supposed to be the main source of information. As their readers had very few opportunities to consult lecturers, the academic books had to be written in simple language understandable for the novice. After the 1990 Higher Education Act enabled private universities, these books shifted to supplementing lectures and helping lecturers deliver systematic instruction. Authors wrote monographic encyclopaedias of law with an assumption that any unclear issues would be discussed with lecturers (Izdebski and Murzyn 1968: 10).

4. The study

4.1 The method

The study aims to examine Polish publications titled "encyclopaedia of law" with a view to identifying their characteristics and investigating the relationship between the two genres discovered: lexicographic and monographic encyclopaedias of law.

The sources for this study include the collections of the National Library of Poland, the University of Warsaw Library, the Library of Cardinal Stefan Wyszyński University in Warsaw and the Main Library of the University of

Siedlce. The National Library of Poland, the primary source of data, is one of the libraries that receive two copies of books, serials and other publications under the Act of November 7, 1996 on Legal Deposit Copies (Cornell 2016: 26). Consequently, it holds nearly all items published in Poland since 1996 and maintains extensive pre-1996 collections. The other three libraries supplemented the sample with volumes that had failed to be deposited with the National Library of Poland, including the 23 volumes of *Wielka encyklopedia prawa*. The completeness of the sample was verified using the meta-catalogue available on the website of the National Library of Poland. This meta-catalogue provides access to catalogues from 651 Polish libraries and their branches (university, public, scientific and specialist)⁴.

Publications were identified using the keywords: "encyklopedia prawa" (encyclopaedia of law), "encyklopedia" (encyclopaedia) and "prawo" (law), or "encyklopedia" and "prawo" (the genitive case of "prawo")⁵. Consequently, the study comprises 78 publications, including 39 lexicographic and 39 monographic encyclopaedias of law. For publications with multiple editions, the analysis includes only one — preferably the most recent volume.

The study is organised into two parts. The first provides general characteristics of the sample: timeframe, size of publications (number of pages), their scope (number of entries), thematic coverage and target audience. The second focuses on analysing the contents and structure of individual publications. This analysis follows the division into lexicographic and monographic volumes, and general-law (discussing the legal system as a whole) and narrow-field ones (discussing specific areas of law). These categories were subdivided into types sharing common features, pointing at characteristics recurring within each genre (Figure 1).

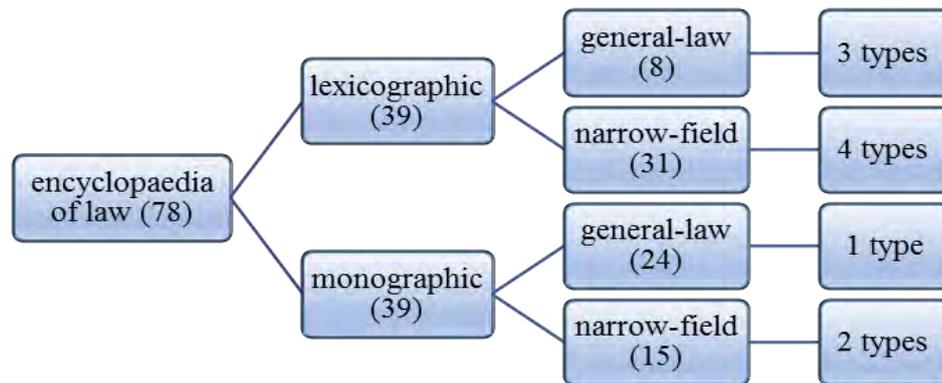


Figure 1: Division of publications in the analysis of contents and structure (the number of publications is given in brackets)

For lexicographic publications, the analysis includes:

- the target group as stated by the authors or implied by the contents,
- the macrostructure, including the arrangement of entries, components of front matter and back matter, and
- the microstructure, including definition type and components of individual entries.

For monographic publications, the analysis focuses on structuring of contents, wording of chapter titles and methods of presenting information. The analysis begins with features common to all monographic works before examining specific categories and types.

4.2 Types of publications and their characteristics

Two types of encyclopaedias of law were identified. One group comprises lexicographic works (39 publications) containing definitions of headwords accompanied by encyclopaedic information in separate alphabetically or thematically arranged entries. A clear separation technique is evident, as each entry is an independent, self-contained unit. However, cross-reference systems in the majority of lexicographic publications highlight strong relations between concepts. The other group — monographic volumes (39 publications) — comprises works structured in chapters and subchapters, covering topics from Encyclopaedia of Law taught primarily in non-legal disciplines.

The sample encompasses works published from 1842 to 2023 (Figure 2). Lexicographic encyclopaedias started appearing in the 1930s (the interwar period) as Polish law systematization after the Partitions of Poland⁶. Many such publications were published in the 21st century: in the 2000s this may be connected with the terminographic boom in Poland (see Łukasik (2015: 88) and Rzepkowska (2018: 66-67)), and in the 2010s and 2020s this results from printing the 23 volumes of *Wielka encyklopedia prawa*. Monographic encyclopaedias of law in Poland date back to the first half of the 19th century (Kłodziński 1842), including early translations from German. In the interwar period, such works appeared in considerable numbers as academic books (9 publications in the 1920s and 2 in the 1930s). Such monographs continued after WWII for non-law students. These students used Encyclopaedia of Law courses to familiarise themselves with the Polish legal system for business and administrative purposes.

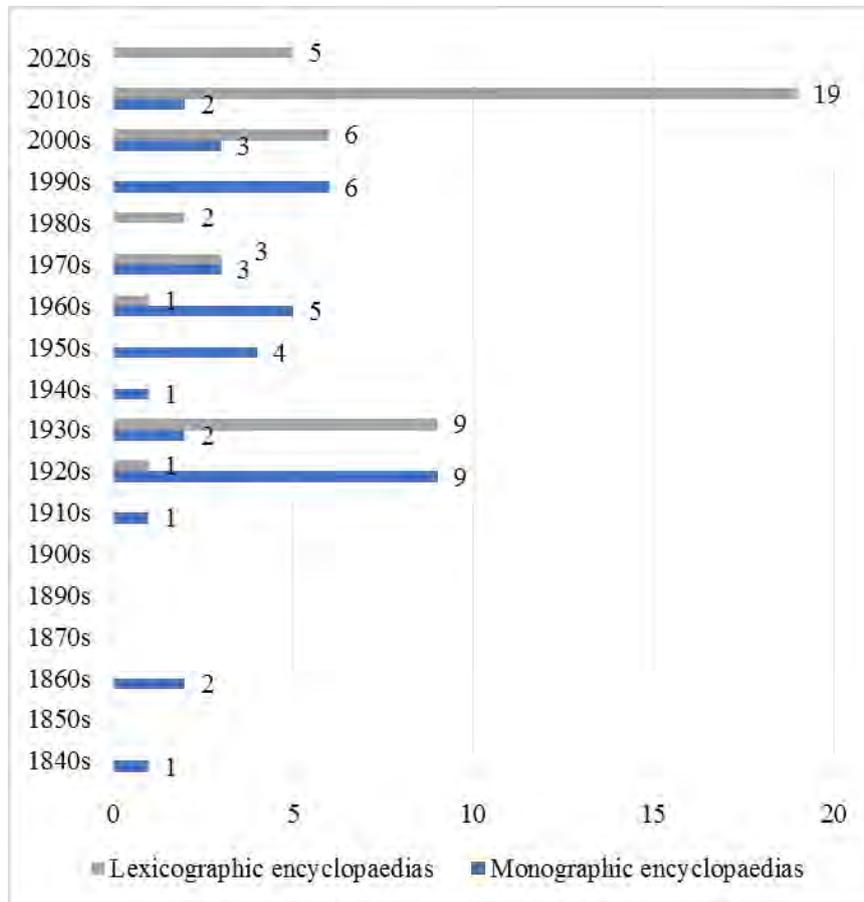


Figure 2: Lexicographic and monographic encyclopaedias by the decade of publication⁷

The majority of monographic encyclopaedias concentrate on Polish law in general (61%). The rest discuss seven fields of law, including administrative and criminal law. The proportions are reversed for lexicographic encyclopaedias: 21% concentrate on general law and 79% on specific fields of law (28 fields in total) (Figure 3). Given the marked differences between general and narrow-field publications in the two groups under analysis, the remaining figures in the paper follow this division.

Monographic and lexicographic encyclopaedias differ in the number of pages (Figures 4 and 5). Monographic encyclopaedias are generally thinner, while lexicographic encyclopaedias are thicker. Narrow-field and general-law lexicographic encyclopaedias are of relatively similar size (from 600 to over 1,000 pages).

The largest publications are the multi-volume narrow-field encyclopaedias from the 1920s and 1930s, for example a 4-volume encyclopaedia with 2,500 pages (Konic 1931, Zoll and Wasilkowski 1936, 1938, 1939). Monographs are much smaller: 54% have under 200 pages and 92% under 400 pages.

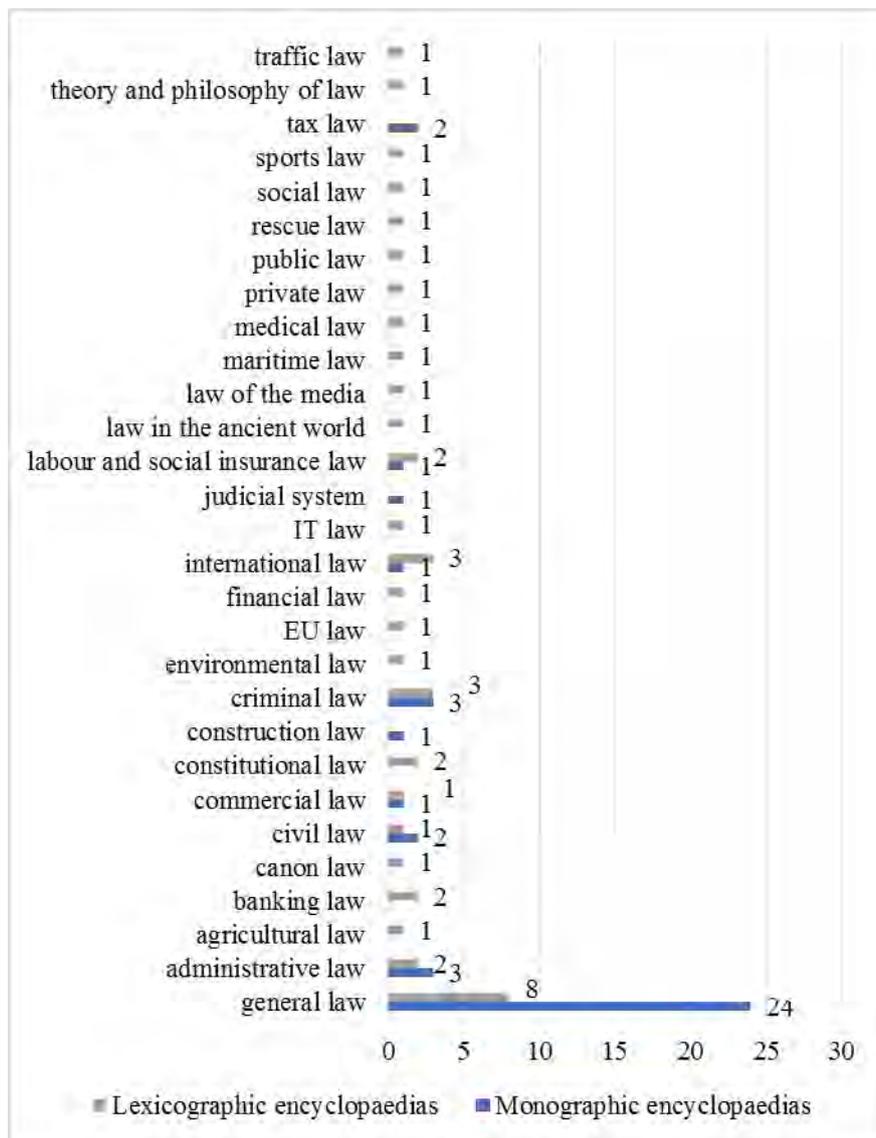


Figure 3: Monographic and lexicographic encyclopaedias by field of law

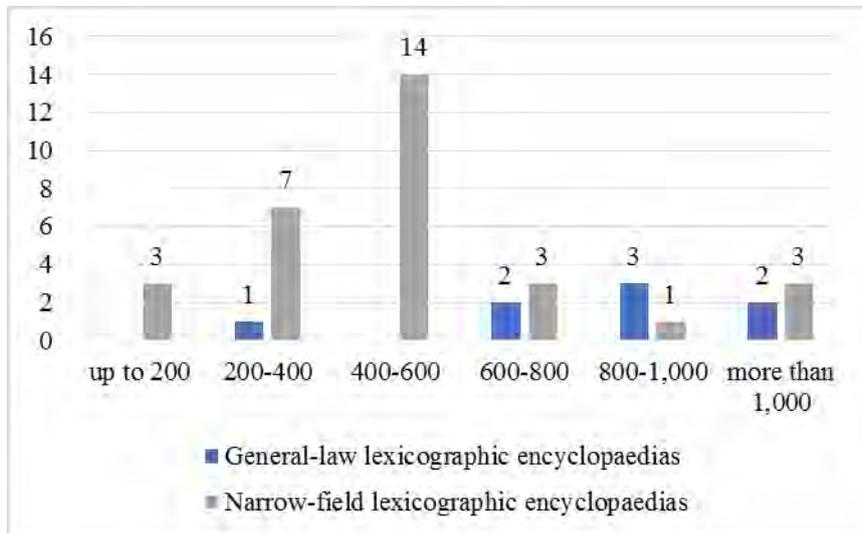


Figure 4: Lexicographic encyclopaedias by number of pages

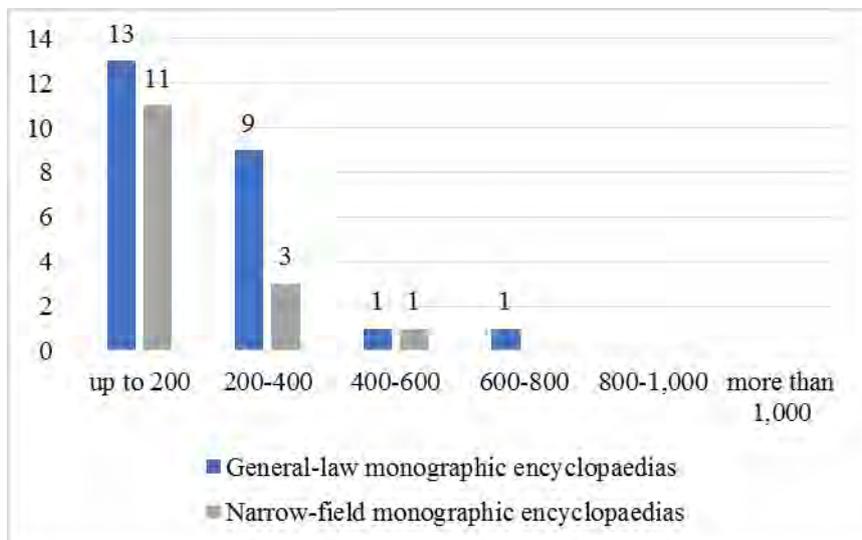


Figure 5: Monographic encyclopaedias by number of pages

The analysis of publications by number of entries applies to lexicographic encyclopaedias only (Figure 6). The largest number of entries is found in general-law encyclopaedias. Over 60% of the narrow-field sample contains between 100 and 500 entries. Page-to-entry ratios imply that entries in narrow-field publica-

tions are extensive⁸ and those in general-law books short, which corresponds to their respective purposes: meeting the needs of professionals and semi-professionals in the former case, and the needs of non-professionals in the latter, respectively.

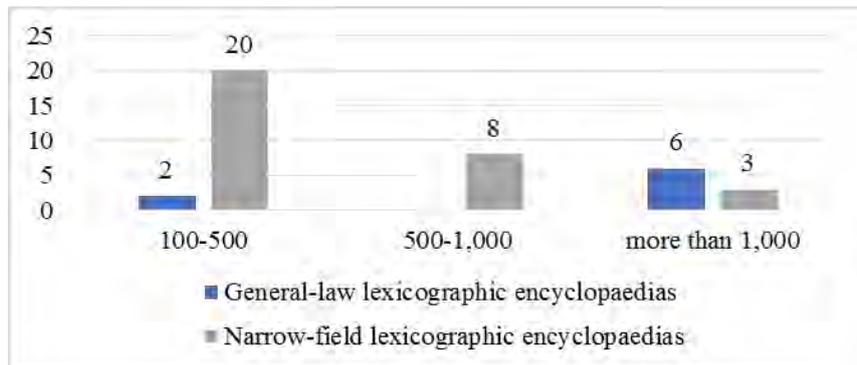


Figure 6: Lexicographic encyclopaedias by number of entries (approximate⁹)

All general-law lexicographic encyclopaedias are addressed to non-professionals (the general public), but professionals (lawyers) and semi-professionals (law students) may benefit from them as well, as the authors note¹⁰. Narrow-field lexicographic publications are often designed as multi-purpose volumes, attempting to meet various users' needs (Figure 7). The majority of general-law monographic volumes are targeted at non-professionals (mainly students of non-legal studies), while the narrow-field ones are mostly addressed to professionals and semi-professionals (Figure 8).

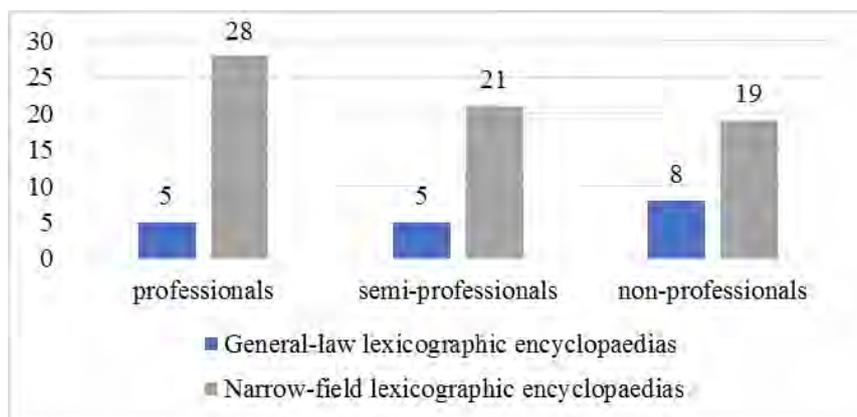


Figure 7: Lexicographic encyclopaedias by target groups

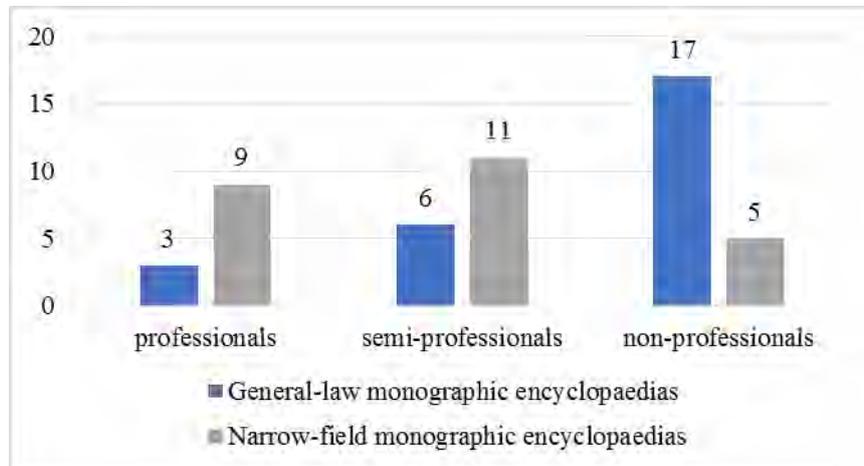


Figure 8: Monographic encyclopaedias by target groups

4.3 Lexicographic encyclopaedias of law: Structure and contents

4.3.1 General-law lexicographic encyclopaedias

Type 1

The first group comprises encyclopaedias featuring a relatively simple macrostructure with an alphabetical order of entries and additional components in the front matter — such as a preface, a list of abbreviations, a list of illustrations, Latin phrases, basic literature and a list of authors with initials (e.g., in Hołyst 2005 and Kalina-Prasznic 2004, 2005) — and in the back matter — such as an index of persons (Kurowski 1960), an afterword (Rybicki 1980) and a list of tables (Kalina-Prasznic 2004). The microstructure comprises short classical definitions by genus and difference ("A is B, that is C"), usually supplemented with encyclopaedic information, typically yielding several entries per page. The 20th-century publications include authors' initials at the end of each entry. Encyclopaedias in this group include volumes by Kurowski (1960), Rybicki (1980), Hołyst (2005) and Kalina-Prasznic (2004, 2005). The typical target group is non-lawyers (engineers, doctors, economists, etc.), law students and lawyers who wish to learn about topics from beyond their specialty in law. The short classical definitions make them appear more like dictionaries than encyclopaedias, but the fact that the volumes present such content as the history of law, names of places as entries, and a list of authors of political and legal doctrines moves them towards the encyclopaedic type. These lexicographic encyclopaedias use cross-referencing to provide broader context for the headwords and save space in individual entries, revealing the systemic nature of law.

Type 2

This type contains two publications selected based on their target group — both are addressed to non-experts in law: Garlicki (1974) and Kerplik et al. (2009). By avoiding discussions about controversial issues that, due to their specificity, would be addressed to lawyers, they present information in a simplified manner that can be easily understood by those without a legal background. The authors of individual entries are identified in both cases: authors' initials appear at the end of each entry in the former and a list of headwords is provided alongside the authors' names in the front matter in the latter. In Garlicki (1974), the entries are alphabetically arranged within eight thematic sections, each devoted to one field of law, creating the impression of several narrow-field encyclopaedias within a single volume. The headwords are listed in the contents found in the front matter. There is also an index of entries in the back matter with page references. In Kerplik et al. (2009), the headwords are in alphabetical order. They are also listed at the beginning of each alphabetical section. The microstructure of entries in this volume is presented with varied visual formatting, with content in different font sizes and styles, sometimes italicised or bold to attract the reader's attention, and includes illustrations where applicable. Each entry is divided into sections — for example, in the case of "abortion": "remember" with practical information about the concept discussed, "legal liability" and "legal basis". The back matter comprises an index of legal terms and the titles of sections in which they can be found.

Type 3

This group comprises only one publication that is notable in terms of its structure: the work by Bielecka-Pieczka et al. (2011). It is a thematic encyclopaedia organising the essence of general law into 16 themes (Figure 9). The title literally means that it is an encyclopaedia of law for everyone, which suggests that the content is designed to meet the needs of the general public. Each theme is developed through several topics (see the theme "employment law" in Figure 9). Neither themes nor content within each theme are alphabetically arranged. The only alphabetical arrangement is used in glossaries that conclude each theme. The glossary entries comprise classical definitions of the headwords, only sometimes supplemented with encyclopaedic information such as the applicability of a given legal concept. Additional context is provided in the sections preceding each glossary. There are elements that are repeated in each theme, such as subchapter titles posed as questions at the beginning and the glossary at the end. The themes are several pages long (over 40 pages in the case of the "employment law" theme).

<i>Encyklopedia prawa dla każdego</i> (Bielecka-Pieczka et al. 2011)	
Example contents of a theme	
Employment law	
– What's employment law?	– Other ways of terminating the contract
– Regulations	– Expiration of employment contract
– Employer and employee	– Collective redundancies
– Employment relationship	– Work certificate
– Civil-law contracts	– Remuneration for work
– Appointment	– Working time
– Selection	– Working time standard
– Nomination	– Days off
– Cooperative contract of employment	– Annual leave
– Appropriate employment contract	– Unpaid leave
– What information the employer cannot ask for?	– Maternity and paternity leave
– Supervising the employee	– Child care leave
– Employer's duties	– Other types of leave
– Rules	– Employment law disputes
– Medical examination	– Labour court
– Mobbing	– Where to look for help?
– Harassment	– Occupational accident
– Termination of employment contract	– Occupational disease
– Mutual agreement	– Accident on the way to or from work
– Termination with notice	– Social insurance and benefits
– Termination without notice	– Glossary
	– Specimens of documents

Figure 9: Excerpt from the table of contents: "Employment law" (Bielecka-Pieczka et al. 2011)

4.3.2 Narrow-field lexicographic encyclopaedias

Type 1

The review of narrow-field encyclopaedias starts with publications from the inter-war period, which are extraordinary works considering their size and their function as a means of unifying law in the Polish territories after the partitions. The time of their publication coincided with the formation of a new legal regime in Poland out of five different civil legal systems in force in its territory before Poland regained independence in 1918 (Konic 1931: "Przedmowa"). All three encyclopaedias in this group — by Cybichowski (1926, 1930), Makowski (1931–1934,

1934–1936, 1936–1937, 1938–1939) and Konic (1931), Zoll and Wasilkowski (1936, 1938, 1939) — are multi-volume publications. In terms of their macrostructure, entries are arranged alphabetically (the first volumes start with the letter A and the last volumes end with Z). The front matter includes a foreword and a list of authors with their initials (in Makowski's encyclopaedia and in Konic, Zoll and Wasilkowski's publications). Thematically, each covers an area of the contemporary legal system: Cybichowski's covers public law, Makowski's covers criminal law and Konic, Zoll and Wasilkowski's covers private law. As for the microstructure, the structure of the entries in the encyclopaedias varies depending on the subject and the author of a particular article. Rather than providing concise definitions, they contain descriptive essays providing broad background on the relevant legal concept. Each entry is a self-contained, usually lengthy, dictionary article, often divided into separate sections (numbered or unnumbered chapters and subchapters), ending with source literature and the name or initials of the author. In one of the encyclopaedias, some entries start with a table of contents depicting the structure of the articles (e.g., "Księgi wieczyste" (Zoll and Wasilkowski 1936: 615) and "Osobistość prawna" (Zoll and Wasilkowski 1938: 1245)). Although the authors do not specify it, the magnitude and degree of detail of these publications made them suitable for lawyers and law students (in Konic, Zoll and Wasilkowski's publications, approximately 250 entries across 2,500 pages, and in Cybichowski's, 416 entries across nearly 1,200 pages).

Type 2

The next group of narrow-field encyclopaedias concentrates on publications addressed to non-legal professionals — namely, experts in the professions that law focuses on. This group includes volumes by Brojewski et al. (1975), Łopuski and Kowalewski (1982), and Klafkowski (1976). What they have in common is a relatively complex macrostructure (including detailed explanations of the microstructure and macrostructure (Łopuski and Kowalewski 1982; Klafkowski 1976) and an index grouping entries thematically in the back matter), alphabetical arrangement of entries and, microstructurally, cross-references to other entries and intensive use of abbreviations. The first volume (Brojewski et al. 1975) is a concise encyclopaedia of employment and social insurance law. Entries (about 110 in total) do not contain classical definitions of headwords but only encyclopaedic information in the context of employment and social insurance. The information is relatively detailed but without legal jargon, making it accessible to non-lawyers. In Łopuski and Kowalewski's volume (1982), the back matter is extensive with a thematic index of headwords (with 18 themes under which the headwords are listed) and a thematically arranged reference list and indices of foreign language equivalents (English, French, German and Russian). This is the only publication in the sample that includes foreign language equivalents. Each entry starts with a classical definition followed by encyclopaedic information supported by illustrations. It ends with a list of equivalents in four languages. Klafkowski's encyclopaedia (1976) is distinctive for its developed front matter,

which comprises a preamble, initial remarks, general abbreviations and acronyms of institutions with their full names, an explanation of the system of references, a list of Latin phrases used in the encyclopaedia and their explanations. In terms of microstructure, there are entries with classical definitions supplemented with encyclopaedic information on the topic and entries containing only encyclopaedic information. Each entry ends with references to legal sources and literature, and the full name of the entry author.

Type 3

This group of encyclopaedias has been grouped together because they were published in an easily updatable paper form using a ring-binder, which accommodates the changing nature of law (these are Pyziół 2001 and Florek 2003). Both alphabetically arranged publications are large (nearly 1,000 and over 600 pages, respectively) and addressed to lawyers, law students, students of related disciplines and those interested in legal issues related to banking (Pyziół 2001) and labour law and social insurance (Florek 2003). The front matter in Pyziół (2001) comprises a list of authors, a page for recording updates, a list of abbreviations, a foreword and a list of entries with page numbers. The entries are relatively long (from 2–3 to even 20 pages; a little over 200 entries across about 1,000 pages), providing substantial encyclopaedic and specialist information in addition to defining headwords, yet without presenting definitions that summarise the concept in question. They do not follow any strict structure. In the encyclopaedia by Florek (2003), the front matter comprises an alphabetical index of entries and a thematic index with page numbers. In addition, there are a user guide and a list of abbreviations. Each entry follows a predefined structure: a concise definition of the headword, its legal basis, a detailed description of the institution and references.

Type 4

*Wielka encyklopedia prawa*¹¹, published between 2014 and 2023, is a collection of 23 volumes with 11,921 entries across 11,000 pages¹². The volumes discuss different areas of law individually, which is why in this paper they are treated as separate publications, not as a single encyclopaedia. The fields discussed in individual volumes include: the laws of antiquity, canon law, EU law, public international law, environmental law, constitutional law, theory and philosophy of law, agricultural and food law, law of the media, medical law, financial law, social law, criminal law, commercial law, rescue law, banking law, public commercial law, substantive law, traffic law, administrative law, law of criminal procedure, IT law and sports law. Each volume has its own editors and follows roughly the same structure. As for the macrostructure, the front matter is well-developed. It comprises an introduction, a list of authors of entries, information about abbreviations used, a list of sources (usually divided into different categories of sources depending on the field of law and applicability) and references to literature. When

examining the microstructure, one can notice that each entry starts with a classical definition and is supplemented with information of an encyclopaedic character. It ends with the author's initials, followed by a reference to recommended literature. What can be found within individual entries is a system of cross-referencing. The entries are of varying length: the shortest only refer readers to other entries, whilst the longest can be a few pages long. The volumes are of varying sizes: from 192 pages (Skoczylas 2018) to 668 pages (Pyć et al. 2021) and from approximately 160 entries (Bagińska 2016) to about 1,000 entries (Bator et al. 2016; Szpor and Grochowski 2021).

4.4 Monographic encyclopaedias of law: Structure and contents

Monographic encyclopaedias of law follow the structure outlined in the tables of contents typically found in the front matter. The information is divided into chapters representing themes which are either general in scope (resulting in fewer chapters) or highly specific (resulting in more numerous chapters). These themes are then divided into related topics which are further subdivided into subtopics (subchapters). The hierarchical structure created in this way serves to reconstruct the system of relations within the relevant domain, resembling that of thematic encyclopaedias.

The back matter typically contains thematic or alphabetical indices, but having these components is not a standard for these publications. Generally, chapters and subchapters take the form of articles of varying length describing the concepts discussed. The titles of chapters are sometimes very descriptive (e.g., Kowalczyk 2002), but in some cases they are limited to the legal term on which the chapter focuses when the chapters are very detailed (e.g., Siuda 1975; Łopatka 1995a, 1995b; Bossowski 1925; Guzy 1967). When this is the case, they may be comparable to the headwords in lexicographic publications on law.

4.4.1 General-law monographic encyclopaedias

Type 1

Monographic general-law encyclopaedias comprise publications intended to define Encyclopaedia of Law as an academic discipline and academic subject¹³. They were frequently written in connection with the course in Encyclopaedia of Law directly for university students (as a source of information) and indirectly for lecturers (as a way to unify the contents of lectures delivered by different scholars). The majority of publications of this type are addressed to non-law students and often take the form of publications for internal university or higher education institutions' needs. They are written with the special needs of those studying particular subjects in mind, providing them with general information about the Polish legal system, for example, in economics (Siuda 1975), management and marketing (Muszyński 2001) or tourism (Kowalczyk 2002). Figures 10 and 11 present excerpts from the tables of contents of selected monographs.

<i>Encyklopedia prawa: część 1 (Łopatka 1995a)</i>
Example contents of a chapter
Chapter XVIII. Sources of law
1. The concept of sources of law
2. Development vs creation of law
3. Enactment
4. Recognition
5. System of sources of law
6. Legislative drafting
7. A collection of legal provisions, a code
8. System of sources of law in Poland

Figure 10: Excerpt from the table of contents (Łopatka 1995a)¹⁴

<i>Encyklopedia prawa dla ekonomistów (Siuda 1975)</i>
Example contents of a chapter
IV. Criminal law
1. Sources of criminal law
2. Effectiveness of criminal law
1. Effectiveness of the criminal law in time
2. Effectiveness of criminal law in space
3. The notion of crime
4. Types of crimes
5. Cause-and-effect relationship
6. Stages of a criminal act
7. Forms of participation in a crime
8. Guilt
9. Circumstances excluding criminal liability
1. Insanity
2. Mistake
3. Self-defence
4. State of necessity
5. Supervisor's order
6. Consent of the injured
7. acting within the limits of one's authority or legal duty
8. Risk of novelty
<i>[there are 22 subchapters more in this chapter]</i>

Figure 11: Excerpt from the table of contents (Siuda 1975)

4.4.2 Narrow-field monographic encyclopaedias

Type 1

Nine of the 15 narrow-field publications date from the interwar period¹⁵. Like lexicographic publications from that time, their purpose was to systematise and ultimately unify law in the new territory of Poland. Their focus on specific areas of law in various aspects, such as criminal, commercial and tax law, makes them publications for lawyers and law students, rather than for non-lawyers (which is often confirmed by the authors in the prefaces). The content of each subchapter is typically in the form of academic essays providing very detailed information on the topic compared to the general-law publications (see Figure 12).

<i>Encyklopedia prawa obowiązującego w Polsce: praca zbiorowa z udziałem profesorów uniwersytetów polskich. Cz. 2, Prawo karne i proces karny wszystkich dziedzin (Bossowski 1925)</i>
Example contents of a chapter
Part I: Criminal law
I. Criminal lawsuit in the area under former Austrian rule and a supplement [<i>at the back of the book</i>]
Sources, issues of acts of law, academic books, jurisdiction and a supplement [<i>at the back of the book</i>]
C. Regulations about punishment
1. Basic penalties
a) Capital punishment
b) Hard imprisonment
c) Imprisonment
d) Custody
e) Fine
f) Loss of rights and entitlements
g) Reprimand from the judge
h) Imprisonment in a separate jail
2. Additional penalties
a) The case of goods, commodities and equipment
b) Dismissal/Deportation
c) Police surveillance
3. Penal consequences
4. Sentence in case of concurrence of crimes
5. Special mitigation and change of penalty
6. Recognising custody as part of sentence
7. Parole
8. Act of pardon
9. Statute of limitations

Figure 12: Excerpt from the table of contents (Bossowski 1925)

Type 2

The publications published after WWII include four volumes from the 1960s (Guzy 1967, 1968; Muszyński 1969; Saulewicz 1967), one from the 1970s (Bejtka 1977) and one from the 21st century (Domagała et al. 2010). The publications from the 1960s provide basic information about selected areas of law for military personnel, the one from the 1970s for site managers and the one from 2010 for public administration.

In this group, the volume by Guzy (1967) is notable in the context of this paper as it can be placed on the boundary between a lexicographic and monographic work. The table of contents suggests that it is another monograph titled "encyclopaedia of law". However, the titles of chapters are all relatively short, terminology-based noun phrases referring to legal concepts and the contents of chapters resemble those of a lexicographic work, in many cases starting with a classical definition that is further completed with encyclopaedic information. All of this leads to the volume being very similar to a thematic encyclopaedia.

5. Findings and conclusions

The analysis of 78 Polish publications titled "encyklopedia prawa" (1842–2023) revealed the existence of two distinct types of works under this title: lexicographic and monographic encyclopaedias. Each constitutes roughly half of the sample. The lexicographic encyclopaedias are characterized by entries arranged alphabetically (e.g., Kurowski 1960, Hołyst 2005) or thematically, the latter quite infrequently (e.g., Bielecka-Pieczka et al. 2011), presenting fragmented, encyclopaedic information. In contrast, the monographs adopt a systematic chapter-based structure, resembling textbooks or academic introductions to law.

The review of library holdings revealed the following general and detailed characteristics of the two genres:

- **Timeframe** — Lexicographic encyclopaedias emerged primarily in the 20th and 21st centuries, while the monographic encyclopaedias appeared earlier, in the 19th century. Interestingly, interwar narrow-field monographic publications (eight volumes in 1923-1929) preceded the lexicographic encyclopaedias (eight volumes in 1931-1939), as if preparing the ground for the latter.
- **Size** — Lexicographic encyclopaedias were generally larger than monographs. Some entries spanned several pages, particularly in narrow-field publications, and resembled chapters divided into sections, like monographic volumes.
- **Coverage and target group** — Lexicographic encyclopaedias typically specialised in specific fields, while monographs covered general Polish law. General-law lexicographic encyclopaedias target non-professionals (non-law students), while narrow-field encyclopaedias serve: lawyers, law stu-

dents and non-legal professionals interested in specific law. Monographic purposes vary by scope: general-law volumes focus on non-professionals and are rarely considered suitable for professionals, while narrow-field volumes target legal professionals.

- Macrostructure (lexicographic volumes) — Encyclopaedias are typically alphabetical, with one thematically arranged exception. They are usually single-volume, except for large interwar narrow-field encyclopaedias. They sometimes feature extensive front and back matter, alphabetical and thematic indices, lists of abbreviations and illustrations, and reference lists. The twenty-three narrow-field volumes in *Wielka encyklopedia prawa* (2014–2023) maintain consistent structural patterns across multiple volumes while addressing different legal domains. Notable are systematisation techniques, including cross-referencing and thematic indices, demonstrating the need to place legal concepts in wider context, particularly in narrow-field works. This illustrates that restricting law descriptions to discrete entries is difficult considering the systemic and culturally bound nature of law.
- Microstructure (lexicographic volumes) — Entries in general-law works are either short, containing classical definitions with little additional information (type 1), or longer entries containing simplified information for non-professionals (types 2 and 3). Entries in narrow-field publications vary in length and detail. On the one hand, there is extensive and often chapter-based organisation of lexicographic entries in narrow-field encyclopaedias, which address specialist information using expert terminology. Others designed for non-legal professionals (e.g., in the marine industry (Łopuski and Kowalewski 1982)) include entries with classical definitions and profession-focused information, sometimes in illustrated form, or contain only encyclopaedic information.
- Structure (monographic volumes) — Monographic volumes are organised into chapters and subchapters as outlined in tables of contents found in the front or back matter. Chapters group related concepts under thematic categories, with their titles being either concise and terminology-based (1 to 3 words) or descriptive. Chapters comprise articles, following the formula of an academic textbook, particularly in general-law volumes. Notably, Guzy (1967), with its definitional content and terminology-based chapter titles, represents a work on the boundary between lexicographic and monographic publications.

These findings suggest that functional criteria may be more significant than formal structural criteria in understanding lexicographic practice, particularly in specialized domains where traditional boundaries between reference genres become blurred. This perspective could inform future research on specialised lexicographic traditions and help develop frameworks for analysing contemporary reference publishing.

An important observation emerging from this study is the correlation between the structure of monographic and lexicographic works. The former, while structured as academic textbooks with chapters and subchapters rather than discrete entries, display several characteristics that align them with thematic encyclopaedic principles. Their hierarchical organization mirrors the systematic arrangement found in ancient encyclopaedias, presenting legal knowledge in a structured manner that reflects the conceptual relationships within the legal domain. The chapter titles in narrow-field monographic volumes often function analogously to headwords in lexicographic works, while the content provides definitional and explanatory information comparable to encyclopaedia entries. However, both genres serve distinct but complementary lexicographic functions. Traditional encyclopaedias primarily fulfil cognition-oriented functions, providing quick access to factual information and definitions for immediate reference needs. Monographic encyclopaedias, while ostensibly designed for systematic learning, also serve knowledge-organizing functions characteristic of thematic encyclopaedias, presenting comprehensive conceptual frameworks that enable users to understand the systematic relationships within legal knowledge.

For legal lexicography specifically, the study reveals the importance of recognizing multiple approaches to organizing and presenting legal knowledge. The coexistence of different arrangements, entry lengths and target audiences within one titular category demonstrates lexicographic adaptability to specialised user needs.

6. Methodological limitations

Whilst this study employs systematic criteria for analysing Polish encyclopaedias of law, the classification of publications into categories necessarily involved interpretative decisions that may contain elements of subjectivity. The determination of target audiences and primary functions relied on authorial statements where available but required inference from content and structure when explicit guidance was absent. To mitigate potential bias, classification decisions were based on multiple criteria including structural features, content presentation and stated purposes. The large sample size (78 publications) helps ensure that individual classification decisions do not significantly affect the overall analysis of trends in Polish legal encyclopaedic publishing.

7. Endnotes

1. For the purpose of this paper, Encyclopaedia of Law when capitalised should be understood as the name of an academic course.
2. Languages for Special Purposes.
3. Encyclopaedia of Law as a subject taught at universities originates from Germany, from where it spread to other countries, including Russia, France, Belgium and Poland (*Encyklopedia prawa ...* 1917: 7). The origins of Encyclopaedia of Law go back to the 16th century and the heyday of the discipline was observed in the 18th and 19th centuries. It was not merely an

academic course but a science and an academic discipline (for a review of famous German lawyers-encyclopaedists see Slobozhaninov 2022).

4. See <https://katalogi.bn.org.pl>.
5. Polish legal reference works include numerous lexicons ("leksykon prawa", particularly the C.H. Beck series, e.g., Bagińska 2011; Barczewski and Kowalska 2019; Bartnik 2013). Other publications use the adjective "encyclopaedic" ("encyklopedyczny") in their titles, identifying them as encyclopaedic dictionaries (Wołodkiewicz 1986; Litewski 1998). These works comprehensively describe legal concepts and could merit separate analysis, but were excluded from this study as the focus is specifically on publications titled "encyclopaedia of law."
6. There were three partitions of Poland between 1772 and 1795 by Austria-Hungary, the German and the Russian Empires resulting in the loss of sovereignty by Poland for 123 years. Poland regained independence on 11 November 1918. The interwar period that followed (1918–1939) was the time of rebuilding the country, shaping its borderlines and economy, unifying the legal system and developing national identity.
7. Each volume of the multi-volume encyclopaedias from the interwar period (for the list of works, see 4.3.2, Type 1) is treated here separately owing to the long gaps between publishing individual volumes. Subsequently in the analysis, only the volumes of *Wielka encyklopedia prawa* are treated as separate publications, as each constitutes a self-contained work on a specific area of law.
8. For instance, there are approximately 250 entries in the four volumes of *Encyklopedia podręczna prawa prywatnego* (Konic 1931; Zoll and Wasilkowski 1936, 1938, 1939), which is the largest publication in the sample. As a result, entries there are, on average, 10 pages long.
9. The number of entries is rarely provided by authors, so the data in the chart is an estimation of the number of terms by the author of this paper.
10. The purpose of publications was determined based on information from the authors in the introductions or preambles. Where there was no such information given, the assessment of the purpose was based on the contents and other details from the authors.
11. The editor-in-chief of all volumes is Hołyst, but individual volumes have their own editors, specialists in the relevant areas of law.
12. Based on information about the encyclopaedia on the publisher's website: (<https://www.tamprawo.org/wielka-encyklopedia-prawa/>).
13. The publications in this group include, alphabetically: Ahrens 1862; Armdts von Arnesberg 1866; *Encyklopedia (...)* 1917; Glaser 1931; Gnela 1998; Izdebski 1948; Izdebski and Murzyn 1968; Kłodziński 1842; Kowalczyk 2002; Krzymuski 1921; Lewandowski 1993, 2003; Łopátka 1993, 1995a, 1995b, 1998; Murzyn 1974; Muszyński 2001; Siuda 1975, 2010; Wiszniewski 1952, 1956, 1958; Żabczyński 1957.
14. The contents of publications translated by the author.
15. The publications in this group include, alphabetically: Bossowski 1925; Głabiński 1926; Makowski 1925; Peretiatkowicz 1923a, 1923b, 1925, 1926; Szczygielski 1938; Wasiutyński 1929.

8. References

- Allaby, M. 1985. *The Oxford Dictionary of Natural History*. Oxford: Oxford University Press.
- Allan, K. 2006. Dictionaries and Encyclopedias: Relationship. Brown, E.K. (Ed.). 2006. *Encyclopedia of Languages and Linguistics*. Second edition: 573-577. Oxford: Elsevier.

- Bagińska, E. (Ed.).** 2011. *Leksykon prawa cywilnego — część ogólna: 100 podstawowych haseł*. Warszawa: C.H. Beck.
- Bajčić, M.** 2017. *New Insights into the Semantics of Legal Concepts and the Legal Dictionary*. Amsterdam/Philadelphia: John Benjamins.
- Barczewski, M. and E. Kowalska (Eds.).** 2019. *Leksykon prawa własności intelektualnej: 100 podstawowych pojęć*. Warszawa: C.H. Beck.
- Barenot, P.-N.** 2014. A View of French Legal Lexicography — Tradition and Change from a Doctrinal Genre to the Modern Era. Mac Aodha, M. (Ed.). 2014. *Legal Lexicography: A Comparative Perspective*: 11-30. London: Routledge.
<https://doi.org/10.4324/9781315592053> [6 April 2025].
- Bartnik, M. (Ed.).** 2013. *Leksykon polskiej procedury karnej: 100 podstawowych pojęć*. Warszawa: C.H. Beck.
- Bergenholtz, H.** 2012. What Is a Dictionary? *Lexikos* 22: 20-30.
- Bergenholtz, H. and S. Nielsen.** 2002. Terms in the Language of Culture-Dependent LSP Dictionaries. *Lexicographica* 18: 5-18.
<https://doi.org/10.1515/9783484604476.5> [6 April 2025]
- Bergenholtz, H. and S. Nielsen.** 2006. Subject-field Components as Integrated Parts of LSP Dictionaries. *Terminology* 12(2): 281-303.
- Bergenholtz, H. and S. Tarp.** 1995. *Manual of Specialised Lexicography*. Amsterdam/Philadelphia: John Benjamins.
- Bezuglova, O. and A. Kuznetsova.** 2016. Tendencies of Legal Terms Representation in English and Russian Monolingual Terminological Dictionaries. *Journal of Language and Literature* 7(3): 137-140.
- Chromá, M.** 2004. *Legal Translation and the Dictionary*. Tübingen: Max Niemeyer.
- Collison, R.L. and W.E. Preece.** 2024. Encyclopaedia. *Encyclopædia Britannica*.
<https://www.britannica.com/topic/encyclopaedia> [6 April 2025]
- Cornell, M.** 2016. *Libraries in Poland*. Warsaw: Biblioteka Narodowa.
<https://ksiegarnia.bn.org.pl/pdf/Libraries%20in%20Poland.pdf> [6 April 2025]
- Crystal, D.** 2010. *The Cambridge Encyclopedia of Language*. Cambridge: Cambridge University Press.
- De Groot, G.-R. and C.J.P. van Laer.** 2005. Bilingual and Multilingual Legal Dictionaries in the European Union: An Updated Bibliography. *Legal Reference Services Quarterly* 30(3): 149-209.
<https://doi.org/10.1080/0270319X.2011.601715> [6 April 2025]
- De Groot, G.-R. and C.J.P. van Laer.** 2006. The Dubious Quality of Legal Dictionaries. *International Journal of Legal Information* 34(1): 65-86.
- De Groot, G.-R. and C.J.P. van Laer.** 2008. The Quality of Legal Dictionaries: An Assessment. *Maastricht Faculty of Law Working Paper* 2008(6). Maastricht: University of Maastricht.
- Frawley, W.** 1981. In Defense of the Dictionary: A Response to Haiman. *Lingua* 55: 53-61.
- Greenberg, D.** 2014. The Challenges of Compiling a Legal Dictionary. Mac Aodha, M. (Ed.). 2014. *Legal Lexicography: A Comparative Perspective*: 59-74. London: Routledge.
<https://doi.org/10.4324/9781315592053> [6 April 2025]
- Haiman, J.** 1980. Dictionaries and Encyclopedias. *Lingua* 50(4): 329-357.
- Haiman, J.** 1982. Dictionaries and Encyclopedias Again. *Lingua* 56(3-4): 353-355.
- Hartmann, R.R.K. and G. James.** 2002. *Dictionary of Lexicography*. London/New York: Routledge.
- Hayaert, V.** 2011. Les lexiques juridiques: une ambition encyclopédique? Groult, M. (Ed.). 2011. *Les Encyclopédies: construction et circulation du savoir de l'Antiquité à Wikipédia*: 313-328. Paris: L'Harmattan.

- Hutton, C.** 2014. Defining Ordinary Words for Mundane Objects: Legal Lexicography, Ordinary Language and the Word Vehicle. Mac Aodha, M. (Ed.). 2014. *Legal Lexicography: A Comparative Perspective*: 177-200. London: Routledge.
<https://doi.org/10.4324/9781315592053> [6 April 2025]
- Litewski, W.** 1998. *Słownik encyklopedyczny prawa rzymskiego*. Kraków: Towarzystwo Autorów i Wydawców Prac Naukowych "Universitas".
- Louw, P.A.** 1999. Portals to Knowledge: CD-ROM Encyclopaedias, with Specific Reference to *Microsoft Encarta 99 Encyclopedia*. *Lexikos* 9: 283-293.
- Lukszyn, J. (Ed.)**. 1998. *Tezaurus terminologii translatorycznej*. Warszawa: Wydawnictwo Naukowe PWN.
- Lukszyn, J. (Ed.)**. 2005. *Języki specjalistyczne. Słownik terminologii przedmiotowej*. Warszawa: UW.
- Łukasik, M.** 2015. Quality of Multilingual Specialised Lexicography in Poland. *Cognitive Studies/ Études Cognitives* 15: 79-95.
- Mattila, H.E.S.** 2016. Legal Vocabulary. Tiersma, P.M. and L.M. Solan (Eds.). 2016. *The Oxford Handbook of Language and Law*: 27-38. Oxford/New York: Oxford University Press.
- McArthur, T.** 1986. *Worlds of Reference. Lexicography, Learning and Language from the Clay Tablet to the Computer*. Cambridge: Cambridge University Press.
- McArthur, T.** 1996. *The Oxford Companion to the English Language*. Oxford/New York: Oxford University Press.
- Michta, T. and K. Mroczyńska.** 2022. *Towards a Dictionary of Legal English Collocations*. Siedlce: Wydawnictwo Naukowe UPH.
- Nielsen, S.** 1994. *The Bilingual LSP Dictionary. Principles and Practice for Legal Language*. Tübingen: Gunter Narr.
- Nielsen, S.** 2008. The Effect of Lexicographical Information Costs on Dictionary Making and Use. *Lexikos* 18: 170-189.
<https://doi.org/10.5788/18-0-483> [6 April 2025]
- Nielsen, S.** 2010. Legal Translation Dictionaries for Learners. Karpova, O. and F. Kartashkova (Eds.). 2010. *New Trends in Lexicography: Ways of Registering and Describing Lexis*: 221-232. Newcastle upon Tyne: Cambridge Scholars.
- Nielsen, S.** 2023. Legal Lexicography and Legal Information Tools. Biel, Ł. and H.J. Kockaert (Eds.). 2023. *Handbook of Terminology: Volume 3. Legal Terminology*: 432-457. Amsterdam/New York: John Benjamins.
- Peters, P.** 2015. Transcending Their Formats: Dictionaries, Encyclopedias and Atlases. *Humanities Australia* 6: 68-81.
- Ptaszynski, M.O.** 2010. Checking Knowledge in Online Encyclopaedias: Towards a Behavioural Approach to Data Accessibility. *Lexikos* 20: 410-424.
- Rzepkowska, A.** 2018. *Interdyscyplinarny słownik terminologiczny*. Siedlce: Wydawnictwo Naukowe Uniwersytetu Przyrodniczo-Humanistycznego w Siedlcach.
- Šarčević, S.** 1988. Bilingual and Multilingual Legal Dictionaries: New Standards for the Future. *Revue générale de droit* 19(4): 961-978.
<https://doi.org/10.7202/1058506ar> [6 April 2025]
- Šarčević, S.** 1989. Conceptual Dictionaries for Translation in the Field of Law. *International Journal of Lexicography* 2(4): 277-293.
<https://doi.org/10.1093/ijl/2.4.277> [6 April 2025]
- Slobozhaninov, O.K.** 2022. Development of the Academic Discipline "Encyclopaedia of Law" in Germany in XVIII-XIX Centuries. *Juridical Journal of Samara University* 8(2): 88-94.

- Szemińska, W.** 2011. Translating Law into a Dictionary. A Terminographic Model. *Research in Language* 9(1): 177-186.
<https://doi.org/10.2478/v10015-011-0012-2> [6 April 2025]
- Szemińska, W.** 2014. Translating Law into Dictionaries, or Why One Dictionary Is Not Enough. Budin, G. and V. Lušický (Eds.). 2014. *Languages for Special Purposes in a Multilingual Transcultural World. Proceedings of the 19th European Symposium on Languages for Special Purposes, 8–10 July 2013*: 118-125. Viena, Austria: University of Viena, Centre for Translation Studies.
- Tarp, S.** 2000. Theoretical Challenges to Practical Specialised Lexicography. *Lexikos* 10: 189-208.
- Wojan, K.** 2013. *Język angielski w polskiej leksykografii. Słowniki przekładowe lingwistyczne i encyklopedyczne, wydane w latach 1782–2012*. Gdańsk: Wydawnictwo Uniwersytetu Gdańskiego.
- Wołodkiewicz, W. (Ed.)**. 1986. *Prawo rzymskie: słownik encyklopedyczny*. Warszawa: Wiedza Powszechna.

9. Lexicographic encyclopaedias of law

- Bagińska, E. (Ed.)**. 2016. *Wielka Encyklopedia Prawa. Vol. 10. Prawo medyczne*. Warszawa: Fundacja "Ubi societas, ibi ius".
- Balandynowicz, A., S. Pikulski and J. Warylewski (Eds.)**. 2018. *Wielka Encyklopedia Prawa. Vol.13. Prawo karne*. Warszawa: Fundacja "Ubi societas, ibi ius".
- Bator, A., J. Zajadło and M. Zirk-Sadowski (Eds.)**. 2016. *Wielka Encyklopedia Prawa. Vol. 7. Teoria i filozofia prawa*. Warszawa: Fundacja "Ubi societas, ibi ius".
- Belka, M. and A. Glapiński (Eds.)**. 2019. *Wielka Encyklopedia Prawa. Vol. 16. Prawo bankowe*. Warszawa: Fundacja "Ubi societas, ibi ius".
- Bielecka-Pieczka, A. et al.** 2011. *Encyklopedia prawa dla każdego*. Poznań: Publicat Wydawnictwo.
- Brodecki, Z. (Ed.)**. 2014. *Wielka Encyklopedia Prawa. Vol. 3. Prawo Unii Europejskiej*. Warszawa: Fundacja "Ubi societas, ibi ius".
- Brojewski, M. et al.** 1975. *Prawo na co dzień: encyklopedia podręczna: prawo pracy i ubezpieczeń społecznych*. Warszawa: Wiedza Powszechna.
- Brzeziński, B., H. Litwińczuk and Z. Ofiarski (Eds.)**. 2017. *Wielka Encyklopedia Prawa. Vol. 11. Prawo finansowe*. Warszawa: Fundacja "Ubi societas, ibi ius".
- Cybuchowski, Z.** 1926. *Encyklopedia podręczna prawa publicznego (konstytucyjnego, administracyjnego, i międzynarodowego) [...] Vol. 1, Abolicja — Państwo*. Warszawa: Biblioteka Polska.
- Cybuchowski, Z.** 1930. *Encyklopedia podręczna prawa publicznego (konstytucyjnego, administracyjnego, i międzynarodowego) [...] Vol. 2, Państwo związkowe — Źródła lecznicze*. Warszawa: Biblioteka Polska.
- Florek, L. (Ed.)**. 2003. *Encyklopedia prawa pracy i ubezpieczeń społecznych: komentarz do instytucji prawnych*. Warszawa: "ABC".
- Garlicki, S. (Ed.)**. 1974. *Prawo na co dzień: encyklopedia podręczna*. Warszawa: Wiedza Powszechna.
- Górski, M. and J. Stelmasiak (Eds.)**. 2014. *Wielka Encyklopedia Prawa. Vol. 5. Prawo ochrony środowiska*. Warszawa: Fundacja "Ubi societas, ibi ius".
- Hauser, R., Z. Niewiadomski and A. Wróbel (Eds.)**. 2020. *Wielka Encyklopedia Prawa. Vol. 20. Prawo administracyjne*. Warszawa: Fundacja "Ubi societas, ibi ius".
- Hofmański, P., P. Kruszyński and B. Sygit (Eds.)**. 2021. *Wielka Encyklopedia Prawa. Vol. 21. Prawo karne procesowe*. Warszawa: Fundacja "Ubi societas, ibi ius".
- Holyst, B. (Ed.)**. 2005. *Wielka encyklopedia prawa*. Warszawa: Wydawnictwo Prawo i Praktyka Gospodarcza.
- Kalina-Prasznic, U. (Ed.)**. 2004. *Encyklopedia prawa*. Warszawa: C.H. Beck.

- Kalina-Prasznic, U. (Ed.).** 2005. *Mała encyklopedia prawa*. Warszawa: C.H. Beck.
- Katner, W.J. (Ed.).** 2018 *Wielka Encyklopedia Prawa. Vol. 14. Prawo handlowe*. Warszawa: Fundacja "Ubi societas, ibi ius".
- Kerplik, J. et al.** 2009. *Prawo: encyklopedia*. Warszawa, Bielsko-Biała: Wydawnictwo Szkolne PWN.
- Klafkowski, A. (Ed.).** 1976. *Encyklopedia prawa międzynarodowego i stosunków międzynarodowych*. Warszawa: Wiedza Powszechna.
- Konic, H. (Ed.).** 1931. *Encyklopedia podręczna prawa prywatnego. T. 1, Agent handlowy — księgi handlowe*. Warszawa: Instytut Wydawniczy "Biblioteka Polska".
- Krześniak, E.J. and A.J. Szwarz (Eds.).** 2023. *Wielka Encyklopedia Prawa. Vol. 24. Prawo sportowe*. Warszawa: Fundacja "Ubi societas, ibi ius".
- Kuniewicz, Z., K. Piasecki and B. Ziemianin (Eds.).** 2019. *Wielka Encyklopedia Prawa. Vol. 18. Prawo cywilne materialne*. Warszawa: Fundacja "Ubi societas, ibi ius".
- Kurowski, L. (Ed.).** 1960. *Mała encyklopedia prawa*. Warszawa: Państwowe Wydawnictwo Naukowe.
- Łabno, A. (Ed.).** 2016. *Wielka Encyklopedia Prawa. Vol. 6. Prawo konstytucyjne*. Warszawa: Fundacja "Ubi societas, ibi ius".
- Leszczyński, G. (Ed.).** 2014. *Wielka Encyklopedia Prawa. Vol. 2. Prawo Kanoniczne*. Warszawa: Fundacja "Ubi societas, ibi ius".
- Łopuski, J. and E. Kowalewski.** 1982. *Encyklopedia podręczna prawa morskiego*. Gdańsk: Wydawnictwo Morskie.
- Makowski, W. (Ed.).** 1931–1934. *Encyklopedia podręczna prawa karnego: (prawo karne materialne i formalne, karne skarbowe i administracyjne; socjologia i psychologia kryminalna; medycyna i psychiatria sądowa; kryminalistyka i więziennictwo). Vol. 1, Aberratio ictus-Ecuador*. Warszawa: Biblioteka Polska.
- Makowski, W. (Ed.).** 1934–1936. *Encyklopedia podręczna prawa karnego: (prawo karne materialne i formalne, karne skarbowe i administracyjne; socjologia i psychologia kryminalna; medycyna i psychiatria sądowa; kryminalistyka i więziennictwo). Vol. 2, Egipt-Międzynarodowe prawo karne*. Warszawa: Biblioteka Polska.
- Makowski, W. (Ed.).** 1936–1937. *Encyklopedia podręczna prawa karnego: (prawo karne materialne i formalne, karne skarbowe i administracyjne; socjologia i psychologia kryminalna; medycyna i psychiatria sądowa; kryminalistyka i więziennictwo). Vol. 3, Międzynarodowe prawo karne-Polska*. Warszawa: Biblioteka Polska.
- Makowski, W. (Ed.).** 1938–1939. *Encyklopedia podręczna prawa karnego: (prawo karne materialne i formalne, karne skarbowe i administracyjne; socjologia i psychologia kryminalna; medycyna i psychiatria sądowa; kryminalistyka i więziennictwo). Vol. 4, Polska*. Warszawa: Biblioteka Polska.
- Prutis, S. (Ed.).** 2015. *Wielka Encyklopedia Prawa. Vol. 8. Prawo rolne i żywnościowe*. Warszawa: Fundacja "Ubi societas, ibi ius".
- Pyć, D., R. Stefański and M. Żylicz (Eds.).** 2021. *Wielka Encyklopedia Prawa. Vol. 19. Prawo komunikacyjne*. Warszawa: Fundacja "Ubi societas, ibi ius".
- Pyzioł, W. (Ed.).** 2001. *Encyklopedia prawa bankowego*. Warszawa : Wydawnictwa Prawnicze PWN.
- Rybicki, Z. (Ed.).** 1980. *Mała encyklopedia prawa*. Warszawa: Państwowe Wydawnictwo Naukowe.
- Sieńczyło-Chlabicz, J. (Ed.).** 2016. *Wielka Encyklopedia Prawa. Vol. 9. Prawo mediów*. Warszawa: Fundacja "Ubi societas, ibi ius".
- Sitek, B. and W. Wołodkiewicz (Eds.).** 2014. *Wielka Encyklopedia Prawa. Vol. 1. Prawa świata antycznego*. Warszawa: Fundacja "Ubi societas, ibi ius".
- Skoczylas, J.J. (Ed.).** 2018 *Wielka Encyklopedia Prawa. Vol. 15. Prawo ratownicze*. Warszawa: Fundacja "Ubi societas, ibi ius".

- Symonides, J. and D. Pyć (Eds.).** 2014. *Wielka Encyklopedia Prawa. Vol. 4. Prawo międzynarodowe publiczne.* Warszawa: Fundacja "Ubi societas, ibi ius".
- Szpor, G. and L. Grochowski (Eds.).** 2021. *Wielka Encyklopedia Prawa. Vol. 22. Prawo informatyczne.* Warszawa: Fundacja "Ubi societas, ibi ius".
- Szurgacz, H. (Ed.).** 2017. *Wielka Encyklopedia Prawa. Vol. 12. Prawo Socjalne.* Warszawa: Fundacja "Ubi societas, ibi ius".
- Walaszek-Pyzioł, A. (Ed.).** 2019. *Wielka Encyklopedia Prawa. Vol. 17. Prawo publiczne gospodarcze.* Warszawa: Fundacja "Ubi societas, ibi ius".
- Zoll, F. and J. Wasilkowski (Eds.).** 1936. *Encyklopedia podręczna prawa prywatnego. Vol. 2, F–K.* Warszawa: Biblioteka Polska.
- Zoll, F. and J. Wasilkowski (Eds.).** 1938. *Encyklopedia podręczna prawa prywatnego. Vol. 3, L–P.* Warszawa: Biblioteka Polska.
- Zoll, F. and J. Wasilkowski (Eds.).** 1939. *Encyklopedia podręczna prawa prywatnego. Vol. 4, P–Ż.* Warszawa: Biblioteka Polska.

10. Monographic encyclopaedias of law

- Ahrens, H.** 1862. *Encyklopedia prawa czyli Rys organiczny nauk prawnych i politycznych na podstawie filozofji prawa osnuty.* [Translated by Julian Głowacki et al.]. Petersburg: s.n.
- Arndts von Arnesberg, K.L.** 1866. *Encyklopedya i methodologia prawa.* [Translated by F.M.]. Kijów: s.n.
- Bejtka, M.** 1977. *Encyklopedia prawa dla kierownika budowy.* Bydgoszcz: Towarzystwo Naukowe Organizacji i Kierownictwa.
- Bossowski, J.J.** 1925. *Encyklopedia prawa obowiązującego w Polsce: praca zbiorowa z udziałem profesorów uniwersytetów polskich. Cz. 2, Prawo karne i proces karny wszystkich dzielnic.* Poznań: Fiszer i Majewski.
- Domagała, M., A. Haładaj and S. Wrzosek (Eds.).** 2010. *Encyklopedia prawa administracyjnego.* Warszawa: C.H. Beck.
- Encyklopedia prawa: wykład z roku 1917 i 1918.* 1917. Warszawa: s.n.
- Głabiński, S.** 1926. *Encyklopedia prawa obowiązującego w Polsce: praca zbiorowa z udziałem profesorów uniwersytetów polskich. Cz. 1 z. 4, Ustrój skarbowy Rzeczypospolitej Polskiej.* Poznań: Fiszer i Majewski Księgarnia Uniwersytecka.
- Glaser, S.** 1931. *Encyklopedia prawa.* Wilno: s.n.
- Gnela, B.** 1998. *Encyklopedia prawa dla ekonomistów.* [Ed. by Cz. Żuławska]. Kraków: Wydawnictwo Akademii Ekonomicznej.
- Guzy, S.** 1967. *Encyklopedia prawa. Cz. 2, Wybrane zagadnienia prawa cywilnego.* Warszawa: Wojskowa Akademia Polityczna.
- Guzy, S.** 1968. *Encyklopedia prawa. Cz. 3, Podstawowe wiadomości z prawa pracy.* Warszawa: Wojskowa Akademia Polityczna.
- Izdebski, Z.** 1948. *Encyklopedia prawa w zarysie.* Katowice: Bratnia Pomoc Studentów Wyższego Studium Nauk Społeczno-Gospodarczych.
- Izdebski, Z. and R. Murzyn.** 1968. *Encyklopedia prawa.* Katowice: Wyższa Szkoła Ekonomiczna.
- Kłodziński, J.** 1842. *Encyklopedya i methodologia obejmująca ogólny rys nauk i wiadomości prawnych.* Warszawa: w Drukarni przy ul. Długiej no 591.
- Kowalczyk, J.** 2002. *Encyklopedia prawa: zarys wykładu w Wyższej Szkole Turystyki i Rekreacji w Warszawie.* Warszawa: Wyższa Szkoła Turystyki i Rekreacji: Estrella.

- Krzymuski, E.** 1921. *Encyklopedia prawa: (teoria prawa): według wykładów*. Kraków: s.n.
- Lewandowski, J.** 1993. *Encyklopedia prawa w zarysie A ... & ... Z*. Warszawa: Szkoła Główna Handlowa.
- Lewandowski, J.** 2003. *Encyklopedia prawa w zarysie*. Warszawa: Szkoła Główna Handlowa. Oficyna Wydawnicza.
- Łopatka, A.** 1993. *Encyklopedia prawa: [skrypt]*. [Cz. 1]. Warszawa: Wydawnictwa PWSBiA.
- Łopatka, A.** 1995a. *Encyklopedia prawa. Cz. 1*. Warszawa: Wydawnictwa Prywatnej Wyższej Szkoły Businessu i Administracji.
- Łopatka, A.** 1995b. *Encyklopedia prawa. Cz. 2*. Warszawa: Wydawnictwa Prywatnej Wyższej Szkoły Businessu i Administracji.
- Łopatka, A.** 1998. *Encyklopedia prawa*. Warszawa: Wyższa Szkoła Handlu i Prawa.
- Makowski, J.** 1925. *Encyklopedia prawa obowiązującego w Polsce: praca zbiorowa z udziałem profesorów uniwersytetów polskich. Cz. 1, z. 3, Zobowiązania międzynarodowe Polski*. Poznań: Fiszer i Majewski.
- Murzyn, R.** 1974. *Encyklopedia prawa w zarysie. Cz. 1*. Katowice: Wyższa Szkoła Ekonomiczna.
- Muszyński, J.** 1969. *Encyklopedia prawa. Cz. 5, Wojskowe prawo karne materialne*. Warszawa: Wojskowa Akademia Polityczna: Wojskowy Instytut Prawniczy.
- Muszyński, J.** 2001. *Encyklopedia prawa*. Warszawa: Wyższa Szkoła Zarządzania i Marketingu.
- Peretiatkowicz, A. (Ed.)**. 1923a. *Encyklopedia prawa obowiązującego w Polsce: praca zbiorowa z udziałem profesorów uniwersytetów polskich. Cz. 1*. Poznań: Fiszer i Majewski.
- Peretiatkowicz, A. (Ed.)**. 1923b. *Encyklopedia prawa obowiązującego w Polsce: praca zbiorowa z udziałem profesorów uniwersytetów polskich. Cz. 2*. Poznań: Fiszer i Majewski.
- Peretiatkowicz, A. (Ed.)**. 1925. *Encyklopedia prawa obowiązującego w Polsce: praca zbiorowa z udziałem profesorów uniwersytetów polskich. Cz. 3, Prawo cywilne*. Poznań: Fiszer i Majewski.
- Peretiatkowicz, A. (Ed.)**. 1926. *Encyklopedia prawa obowiązującego w Polsce: praca zbiorowa z udziałem profesorów uniwersytetów polskich. Cz. 5*. Poznań: Fiszer i Majewski.
- Saulewicz, A.** 1967. *Encyklopedia prawa. Cz. 1, Zagadnienia wojskowego prawa administracyjnego*. Warszawa: Wojskowa Akademia Polityczna im. F. Dzierżyńskiego.
- Siuda, W.** 1975. *Encyklopedia prawa dla ekonomistów*. Warszawa: Państwowe Wydawnictwo Naukowe.
- Siuda, W.** 2010. *Encyklopedia prawa*. Poznań: Wydawnictwo Naukowe Contact.
- Szczygielski, A.D. (Ed.)**. 1938. *Encyklopedia praktyki prawniczej. Część 1, Prawo handlowe*. Warszawa: Wydawnictwo "Prasa Prawnicza".
- Wasiutyński, B.** 1929. *Encyklopedia prawa obowiązującego w Polsce: praca zbiorowa z udziałem profesorów uniwersytetów polskich. Cz. 1 z. 2, Ustrój władz administracyjnych, państwowych i samorządowych*. Poznań: Fiszer i Majewski Księgarnia Uniwersytecka.
- Wiszniewski, J.** 1952. *Encyklopedia prawa: wykład. Z. 1*. Warszawa: Szkoła Główna Planowania i Statystyki.
- Wiszniewski, J.** 1956. *Encyklopedia prawa. Cz. 1*. Warszawa: Oddział Wydawniczy Zaoczných Studiów Ekonomicznych.
- Wiszniewski, J.** 1958. *Encyklopedia prawa*. Warszawa: Oddział Wydawniczy Zaoczných Studiów Ekonomicznych.
- Żabczyński, A.** 1957. *Encyklopedia prawa: przewodnik metodyczny*. Warszawa: Oddział Wydawniczy ZSE.

Noms de Marques dans la 9^e Édition du *Dictionnaire de l'Académie Française*

Jan Holeš, *Département des Langues Romanes, Faculté des Lettres,
Université d'Ostrava, République tchèque*
(jan.holes@osu.cz)
(<https://orcid.org/0000-0003-2270-2073>)

et

Zuzana Honová, *Département des Langues Romanes,
Faculté des Lettres, Université d'Ostrava, République tchèque*
(zuzana.honova@osu.cz)
(<https://orcid.org/0000-0002-8973-2288>)

Résumé: L'article contient une analyse des mots provenant des noms de marques dans la 9^e édition du *Dictionnaire de l'Académie française*. L'échantillon comporte 88 mots, dont la plupart proviennent de noms propres (le plus souvent des noms de famille et des noms de lieux), et de noms communs renvoyant à diverses caractéristiques et fonctions de produits. Ces mots sont formés à l'aide de divers procédés morphosyntaxiques, tels que la composition, la dérivation savante et l'abréviation. L'échantillon contient des emprunts fréquents à des langues européennes et notamment à l'anglais. Du point de vue sémantique, ces mots désignent principalement (a) des appareils, des équipements, des machines, des outils, des moyens de transport, des produits alimentaires, des jeux, et (b) des matériaux, des substances, des produits pharmaceutiques, des fibres synthétiques et des matières plastiques. Une brève analyse comparant la présence de mots issus de noms commerciaux dans les dictionnaires *Larousse* et *Le Robert* accessibles en ligne révèle une certaine instabilité de leur traitement lexicographique.

Mots-clés: LEXICALISATION, DEONOMASTIQUE, NOM PROPRE, NOM DE MARQUE, LANGUE FRANÇAISE, *DICTIONNAIRE DE L'ACADEMIE FRANÇAISE*, *LAROUSSE*, *LE ROBERT*

Abstract: *Brand names in the 9th edition of the Dictionnaire de l'Académie française.* The article contains an analysis of the words labelled as brand names in the 9th edition of the *Dictionnaire de l'Académie française*. The sample contains 88 items, out of which most are based on proper names, usually family and place names, and common nouns referring to various features and functions of products. These words are formed using a range of morphosyntactic processes, such as composition, scholarly derivation and abbreviation. The sample includes frequent borrowings from various European languages and particularly from English. From the semantic point of view, these words mainly designate (a) devices, equipment, machines, tools, means of transport, food products, and games, and (b) materials, substances, pharmaceutical products, synthetic fibres, and plastics. A brief analysis comparing the presence of words originating from brand names in

dictionaries *Larousse* and *Le Robert* reveals a certain instability of their lexicographic treatment.

Keywords: LEXICALIZATION, DEONOMASTICS, PROPER NAME, BRAND NAME, FRENCH LANGUAGE, *DICTIONNAIRE DE L'ACADEMIE FRANÇAISE*, *LAROUSSE*, *LE ROBERT*

1. La nature des noms de marques

Les marques représentent un enjeu commercial, financier, juridique, social et communicatif. Pour Petit (2006: 703), le nom de marque est une unité «tout à fait atypique dans le paysage lexical», parce «qu'il ne possède pleinement les propriétés ni du nom propre, ni du nom commun, ni du terme technique». Cet «hybride sémiotique» (Petit 2000) a longtemps été négligé dans la pratique lexicologique et lexicographique, bien que le nom de marque présente d'importants dilemmes lors de l'élaboration d'un dictionnaire.

Les noms de marques appartiennent à la catégorie assez large des ergonymes, c.-à-d. les noms propres de «réalisations ou de découvertes humaines ayant une réalité matérielle» (Leroy 2004: 34), éventuellement des chrématonymes, terme qui a, selon *International Council of Onomastic Science* (s.d.), le même sens dans plusieurs langues. Sjöblom (2016: 453) propose la catégorie assez générale des *noms commerciaux*, comprenant des noms d'entreprises et des noms de produits, dont la fonction est «d'orienter les choix des consommateurs et des investisseurs et dont l'utilisation poursuit des objectifs économiques» et qui ont, généralement, un propriétaire juridique définissant les modalités de son utilisation¹. Dans certains pays slaves, on rencontre les termes *logonymes* désignant les entreprises et les commerces (Krško 2022) et *marconymes* pour les producteurs, les prestataires de services, les vendeurs et les produits destinés à un usage commercial (Galkowski 2022).

La lexicalisation des noms propres constitue un sujet étudié depuis longtemps: pensons à Nyrop (1913) qui a dédié une cinquantaine de pages à la lexicalisation des noms de personnes, de nationalités et de lieux dans le 4e tome de la *Grammaire historique de la langue française*, à Migliorini (1927) qui a consacré tout un volume à la lexicalisation des noms de personnes dans les langues romanes, à Marouzeau (1950: 159-180) qui parle de la *communisation* des noms propres, etc. Büchi (1991: 139) désigne les mots dérivés d'un nom propre comme mots *déonomastiques* et les bases de la dérivation comme des *éponymes*, en reprenant les deux termes de La Stella (1984). Les mots communs provenant des noms propres dans les dictionnaires français ont été étudiés, entre autres, par Lapierre (1989), Schweickard (1989) et Fontant (1998). Le phénomène de la lexicalisation a été étudié par Altmanova (2013), par de nombreux contributeurs du volume édité par Altmanova et Le Tallec (2019) et, récemment, par Chovancová, Mešková et Krafčíková (2022). Vaxelaire (2005) énumère les types de noms propres qui peuvent apparaître dans les dictionnaires de langue, y compris les noms de produits. Dans le domaine de la lexicographie pratique, il convient de mentionner l'ouvrage de Galey (1997) qui explique l'origine et l'histoire de plus de

600 marques, et surtout le *Dictionnaire de noms de marques courants* de Galisson et André (1998) qui s'inscrit dans le domaine de la lexicographie pédagogique. Les noms de marques sont en effet très courants dans le langage de tous les jours, mais leur opacité et leur charge culturelle empêchent souvent la compréhension pour des locuteurs non natifs. Ces noms relèvent donc de la lexiculture, concept développé notamment par Vaxelaire (2019) et Tonti (2020).

Le *Code de la propriété intellectuelle* définit une marque comme «un signe servant à distinguer les produits ou services d'une personne physique ou morale de ceux d'autres personnes physiques ou morales» (République française 2024: Art. L711-1). Malgré cette définition assez large, certaines marques ne peuvent être enregistrées, par exemple «une marque composée exclusivement d'éléments ou d'indications devenus usuels dans le langage courant [...] ; une marque contraire à l'ordre public ou dont l'usage est légalement interdit; une marque de nature à tromper le public, notamment sur la nature, la qualité ou la provenance géographique du produit ou du service» (Art. L711-2). Selon l'*Institut national de la propriété industrielle* (2024a, ci-après l'*INPI*), les marques verbales, qui seules nous intéressent dans le présent article², sont composées «exclusivement de mots ou de lettres, chiffres ou autres caractères typographiques standards ou d'une combinaison de ceux-ci».

Les noms propres partagent avec les noms communs les fonctions définies par Jakobson (1960) et certaines de ces fonctions sont présentes dans les noms de marques. Ce sont surtout les fonctions expressive et poétique qui sont les plus observables. On les retrouve, par exemple, dans les noms de marques qui consistent dans des jeux de mots ou dans ceux qui cherchent à créer l'idée de quelque chose d'exceptionnel, notamment grâce à l'emploi de noms de personnages mythologiques. Knappová (1992) mentionne la fonction d'*individualisation*, qui semble primordiale pour les marques, et dont le rôle est d'identifier un produit, un service ou l'organisation qui les fabrique ou les fournit et de les différencier de produits, services ou organisations similaires. L'auteure ajoute la fonction *connotative* des noms propres, c'est-à-dire leur capacité à renvoyer aux objets du monde réel (p. ex. aux régions de production d'un produit), et la fonction *descriptive* qui associe le nom de marque à certains aspects du produit. En parlant des noms de marques, nous pouvons également mentionner leur rôle juridique, puisqu'ils empêchent l'utilisation du nom par les concurrents, ainsi que leur rôle psychologique et commercial, car ils servent à influencer le consommateur en éveillant des associations positives en lien avec le producteur ou le produit.

Comme tous les mots, les noms de marques sont soumis à la variation linguistique. Ils sont utilisés soit côte à côte avec des noms communs, soit en concurrence avec d'autres noms de marques. La variation peut être plus marquée que dans les mots communs, étant donné que certains noms de marques ne sont répandus que régionalement et que leur utilisation dans certains contextes se heurte à des problèmes juridiques. Ainsi, le *polystyrène* est désigné par des noms de marques différents en Suisse (*Sagex*) et en Belgique (*Frigolite*). La variation

contextuelle est attestée par la concurrence des noms de marques et des termes chimiques en pharmacie (*paracétamol* vs. *Efféralgan*, *Dafalgan*, *Tylenol*, etc.).

Leroy (2004: 49) parle du problème des « faux noms propres », c'est-à-dire des noms qui, « même lorsqu'ils ont rejoint le lexique courant, restent protégés juridiquement et sont écartés à cause de ces problèmes juridico-publicitaires ». Petit (2006: 702) remarque d'ailleurs que les lexicographes indiquent plus ou moins systématiquement qu'il s'agit de noms déposés, afin de ne pas assimiler ces mots totalement aux noms communs et de protéger juridiquement l'éditeur. Le même soin les amène quelquefois à ajouter le symbole ® et la mention des propriétaires des marques.

Altmanova (2013: 37) explique les raisons de la lexicalisation du nom de marque qui peut avoir lieu soit « lorsque les produits ou les services [...] dominent le marché », soit « lorsque son nom est plus simple que le nom générique correspondant ». En lien avec la lexicalisation, Botton et Cegarra (1994: 146) utilisent le terme de *dégénérescence* qui se produit en trois étapes: (a) la personnalisation (la marque commence à être utilisée comme un nom), (b) la banalisation (le nom désigne une catégorie de produits), et (c) la lexicalisation (son entrée dans le dictionnaire). Le résultat de la dernière étape constitue le sujet de cet article. Son objectif est de passer en revue les matrices de création des noms de marques, de décrire leur prise en compte dans le *Dictionnaire de l'Académie française* (ci-après le *DAF*) et de comparer leur traitement dans quelques autres dictionnaires français.

2. Corpus: la 9^e édition numérisée du *DAF*

Le *DAF*, selon Quemada (1997: I), constitue « ouvrage emblématique, instrument de politique de la langue exécuté à la demande et avec le soutien de l'État [...] qui fait exception dans le paysage lexicographique mondial ». La 9^e édition du *DAF*, qui a été initiée dans les années quatre-vingt du XX^e siècle, vient d'être achevée en 2024. Wooldridge (1998: 310) remarque, à propos de l'importance du projet de la numérisation du *DAF*, que ce dictionnaire constitue « une série unique et homogène de descriptions [...] de différents états du français classique et moderne ». Le procès de la numérisation elle-même, y compris ses enjeux techniques et méthodologiques, a été décrit par Catach (2023).

Le portail du *DAF* contient les versions numérisées de toutes les éditions et offre des fonctionnalités qui s'adaptent aux besoins du XXI^e siècle. Pruvost (2021: 106) remarque que « aucun dictionnaire au monde ne dispose de cette possibilité consistant à suivre un mot [...] sur neuf éditions successives, par un simple clic en haut de l'écran ». La nomenclature a considérablement augmenté, proposant « plus de 53 000 entrées, dont 21 000 entrées nouvelles par rapport à la 8^e édition » (Académie française s.d.). De plus, une rubrique étymologique a été ajoutée systématiquement à toutes les entrées. Enfin et surtout, une fonction de recherche avancée permet d'effectuer des recherches en se fondant sur divers critères comme la catégorie grammaticale, la chronologie, le domaine ou l'étymologie.

En sélectionnant dans la recherche avancée le critère «noms et marques déposés» figurant (étonnamment) parmi les choix possibles de la recherche dans l'étymologie par langue, nous obtenons 96 entrées. De la liste des résultats obtenus, nous avons éliminé les composants des noms de marques qui figurent dans le *DAF* comme entrées indépendantes (p. ex. *box*, *cocotte*, *minute*) et les mots *robert* et *tabloïd*, qui représentent des cas de glissements sémantiques plus radicaux³, pour arriver à une liste des 88 entrées qui figurent dans le Tableau 1. Un mot de notre échantillon apparaît en deux variantes: *yo-yo* et *yoyo*. Tous les mots sont des substantif⁴.

D'un point de vue diachronique, la majorité écrasante des «noms et marques déposés» datent, pour des raisons logiques, du XX^e siècle. Seuls 10 mots datent du XIX^e siècle⁵, qui ont tous été ajoutés à la 9^e édition et dont aucun n'était intégré dans la 8^e édition datant de 1935. Les mots *minerve* et *tartan* ont été intégrés dans le *DAF* depuis la 6^e édition de 1835, mais ces entrées ne concernaient que les acceptions appellatives du mot.

La microstructure d'une entrée du *DAF* est simple (voir la Figure 1). Le mot vedette, précédé d'un astérisque s'il est nouvellement ajouté, y figure en majuscules. La catégorie grammaticale, la définition et l'étymologie sont systématiquement données pour toutes les entrées à partir de la 9^e édition du *DAF*. Les liens hypertexte dirigent les usagers vers des sites externes, parmi lesquels *FranceTerme*. D'autres rubriques apparaissent rarement. Nous trouvons par exemple une remarque sur l'orthographe de *mûroise* signalant que le mot peut s'écrire *mûroise* ou *muroise*, selon les rectifications orthographiques de 1990, une remarque grammaticale concernant la formation du pluriel (*tergals*), ou bien une remarque phonétique dans *skai* signalant que le groupe *ai* se prononce *aille*. Rares sont les notes stylistiques indiquant, par exemple, le caractère obsolète de *gramophone* ou familier de *zip*.

* **VEDETTE** (*prononciation*) **note grammaticale** (nom déposé).

Étymologie: Siècle. Étymon, éventuellement évolution sémantique.

DOMAINE. Définition. Exemples.

- Locutions figurées ou familières.
 - Voir aussi
 - [Terminologie (FranceTerme)]:
 - (Domaine)

Liens externes

- [Francophonie (Office québécois de la langue française)] (Pays)
- [Francophonie (Base de données lexicographiques panfrancophone)] (Pays)

Figure 1: Article prototypique de nom de marque dans le *DAF*

3. Analyse des mots issus de noms de marques dans le DAF

3.1 Motivation des noms de marques

Botton et Cegarra (1994: 85-90) énumèrent plusieurs motivations des noms de marques. Entre autres, le nom de leur créateur ou du fondateur de l'entreprise (*Renault*), la raison sociale (*Sopalin* de *Société de Papier Linge*), les noms décrivant le produit (*M. Propre*), leur composition (*Coca-Cola*), leur origine géographique, les références à divers personnages (parfums *Hermès*) et les créations sans significations.

Les noms propres sont très fréquemment utilisés dans la création de noms de marques. Surtout les patronymes qui servaient dans le passé pour dénommer les entreprises selon leurs fondateurs et leurs propriétaires. Ces noms parfois opaques sont dépourvus du lien intrinsèque entre le nom et le produit. Les mots de notre échantillon⁶ proviennent par exemple de noms propres suivants:

- des noms de famille: *bakélite* (de l'inventeur, L.-H. Baekeland), *kalachnikov* (de l'inventeur M. Kalachnikov), *ripolin* (de l'inventeur C. Riep), *sandow* (selon E. Sandow, culturiste allemand), *rimmel* (selon le parfumeur E. Rimmel);
- des toponymes: *goménol* (localité en Nouvelle-Calédonie), *duralumin* (de *Düren*, ville de Westphalie, et d'*aluminium*). Parfois les noms sont tirés de toponymes classiques: *rhodoïd* (du latin *Rhodanus* 'le Rhône' et d'*-oïd*, de *celluloïd*), *rhoovyl* (du même hydronyme et de *-vyl*, tiré de *vinyle*);
- des noms d'entreprises: *semtex* (de *Semtín Glassworks*, usine tchèque localisée à Semtín, actuellement un arrondissement de la ville de Pardubice, et de l'anglais *explosive*), *delco* (initiales de *Dayton Engineering Laboratories Co*);
- des mythonymes: *mélusine* (la fée Mélusine), *minerve* (la déesse Minerve);
- des prénoms: *bénédictine* (tirée d'un adjectif qui fait référence à Saint-Benoît);
- des ethnonymes: *esquimau* (le nom des habitants des terres arctiques);
- de chrématonymes: *taser* (acronyme du titre d'un roman d'aventures américain de 1911, *Tom Swift and Electric Rifle*).

Les noms communs représentent la deuxième source. Ils ont le pouvoir d'évoquer, surtout pour un usager non initié, certains aspects du produit, sa fonction, sa composition, son goût et autres caractéristiques. Selon Lalanne-Gobet (2009: 32), cette évocation peut devenir encombrante dans le cas de l'extension future de la gamme des produits, empêchant, par exemple, d'associer la marque *Orangina* aux boissons fabriquées à partir de fruits autres que des oranges. Dans notre échantillon, le nom évoque:

- diverses fonctions: *climatiseur* (de *climatiser*), *dictaphone* (de *dicter* et du grec *-phone* 'voix'), *frigidaire* (du latin *frigidarium* 'chambre froide'), *gramophone* (du grec *gramma* 'inscription' et *phone* 'voix'), *mobylette* (de *mobile* et de *bicy-*

clette), *pédalo* (dérivé de *pédale*), *photomaton* (de *photographique* et d'*automatique*), *moulinette* (de *moulinet*), *thermos* (du grec *thermos* 'chaud'), etc.;

- la composition: *fibrociment* (de *fibres* et de *ciment*), *mercurochrome* (de *mercure* et du grec *-chrome* 'couleur'), *penthotal* (de *penthiobarbital*), etc.;
- d'autres aspects: *durite* (probablement de *dur*), *cocotte-minute* (de *coquasse*, *cocasse*, altération de *coquemar* et *minute*, soulignant la rapidité de ce mode de cuisson), *nylon* (de *vinyl* et de *coton*), *vespa* (du latin *vespa* 'guêpe' en allusion au bruit du scooter), etc.

3.2 Analyse formelle des noms de marques

Du point de vue formelle, nous pouvons esquisser une classification selon la matrice de création morphosyntaxique ou lexicale. Les noms de marques répertoriés dans le *DAF* sont créés par exemple comme:

- des composés hybrides, p. ex. franco-grecs, comme *cinémascope* (de *cinéma* et du grec *scopein* 'observer') et *dictaphone* (de *dicter* et du grec *-phone*), gréco-anglais, comme *deltaplane* (du grec *delta* et de *plane*, par analogie avec *aéroplane*) et *monotype* (du grec *mono-* et de l'anglais *linotype*), anglo-grecs, comme *cellophane* (de l'anglais *cellulose* et du grec *phainein* 'apparaître'), *celluloïd* (de l'anglais *cellulose* et du grec *eidos* 'aspect');
- des mots-valises: *bureautique* (de *bureau* et d'*informatique*), *mobylette* (de *mobile* et de *bicyclette*), *nylon* (de *vinyl* et de *coton*), *photomaton* (de *photographique* et d'*automatique*), *mûroise* (de *mûre* et de *framboise*), *ronéo* (de *rotary* et de *néo*), *thyatron* (du grec *thura* 'porte' et de *-tron* d'*électron*), *velcro* (de *vel-* de *velours* et de *-cro* de *crochet*);
- des syntagmes figés: *linotype* (de *line of types* 'ligne de caractères'), *parabellum* (de l'adage latin *si vis pacem, para bellum*), *post-it* (impératif anglais);
- des dérivés et des composés grecs: *klaxon* (du grec *klazein* 'retentir bruyamment'), *pyrex* (du grec *pur* 'feu'), *cataphote* (du grec *cata-* et de *phos* 'lumière'), *crésyl* (dérivé savant de *créosote*), *galalithe* (de *gala-* et de *lithos* 'pierre');
- des dérivés et des composés latins: *frigorifère* (de *frigidarium*), *sonotone* (de *sonus* et de *tonus*);
- des dérivés français: *pédalo* (de *pédale*), *sanisette* (de *sanitaire*), *solderie* (de *solde*), *climatiseur* (de *climatiser*), *moulinette* (de *moulin*), *turbulette* (de *turbulent*);
- des composés français: *cocotte-minute*, *placoplâtre* (*plaque* et *plâtre*);
- des mots abrégés: *solex* (de *Vélosolex*), *box* (de *box-calf*), *invar* (d'*invariable*), *meccano* (du français *mécanique*), *teflon* (de *(poly)te(tra)fl(uorethylene)*), acronymes: *delco*, *taser*, cités plus haut;
- le résultat d'autres procédés: *ozalide* (anagramme de *diazol*, substance utilisée pendant la fabrication du papier);
- d'une onomatopée: *zip*.

- *ex nihilo*: *dacron* dont l'étymologie n'est pas spécifiée dans le *DAF*, mais qui, selon certains, est un mot inventé sur le modèle du *nylon*, forgé par E.I. du Pont de Nemours & Co. (Douglas Harper 2024).

Du point de vue de la néologie externe, les noms de marques sont souvent des emprunts dus aux efforts d'utiliser des mots à la mode, provenant d'une langue en vogue, comme l'anglo-américain, ou résultant de l'origine géographiques du fabricant. Ainsi, notre échantillon contient des emprunts à l'italien (*pianola*, *vespa*), à l'anglais (*caddie*, *liberty*, *post-it*, *ruflette*, *scotch*, *scrabble*, *taser*), à l'allemand (*plexiglas*, du latin *plexi-* et de l'allemand *Glas* 'verre'). L'étymologie est obscure pour *yoyo*/*yo-yo*, mais le mot est peut-être issu d'une langue des îles Philippines d'où provient ce jeu (ATILF — CNRS et Université de Lorraine s.d.). Pour certains noms, comme *skai*, une marque désignant un cuir artificiel, créée et déposée par le groupe allemand Konrad Hornschuch AG, nous n'avons pas pu établir une étymologie fiable, le *DAF* se contentant de remarquer qu'il s'agit du nom d'une marque.

3.3 Analyse sémantique des noms de marques

Vu le caractère des mots de notre échantillon qui désignent surtout des objets et des matières assez spécifiques, la polysémie est rare. Ceci ne veut pas dire que la polysémie n'existe pas du tout, mais elle résulte d'une divergence sémantique qui n'a pas trait au nom de marque. Par exemple, des trois acceptions d'*esquimau*⁷, une seule — 'glace enrobée de chocolat et fixée à un bâtonnet' — concerne le nom de marque. Les deux autres acceptions existaient bien avant le dépôt de la marque: (a) 'ensemble des dialectes parlés par les Esquimaux' et (b) 'vêtement d'enfant d'un seul tenant'. Le mot *triplex* désigne soit un 'appartement réparti sur trois étages', soit un 'verre de sécurité constitué de deux couches de verre entre lesquelles est inséré un film fait de celluloïd ou d'une autre matière plastique'. Cette polysémie résulte sans aucun doute d'une double formation indépendante, le premier terme étant formé sur le modèle de *duplex* et le second provenant d'un nom de marque soulignant la composition en trois couches de son produit⁸. Les deux ont néanmoins en commun leur source étymologique, l'adjectif latin *triplex*. *Polaroid* représente une exception unique, possédant deux acceptions qui résultent d'une métonymie et qui sont associées au nom de marque: (a) 'feuille transparente de résine synthétique polarisant la lumière qui la traverse' et (b) 'appareil photographique non numérique à développement instantané'.

La polysémie peut résulter aussi de l'usage métonymique du nom de l'entreprise pour ses produits. Fèvre-Pernet et Roché (2005) proposent de distinguer les *noms de marques* et les *noms de produits* dans l'onomastique commerciale, parce que les deux catégories manifestent des différences sur le plan pragmatique, référentiel, syntaxique et morphologique. Ainsi, dans notre échantillon, le mot

velux désigne une 'fenêtre à un seul châssis percée dans la pente d'un toit' ainsi que la société danoise Velux produisant des fenêtres de toit.

Du point de vue des champs sémantiques, nous distinguons en particulier deux groupes de référents:

- (a) une cinquantaine de mots désignant des appareils, des machines, des produits, des procédés et des outils: *bureautique, caddie, cataphote, cinémascope, cinématographe, climatiseur⁹, cocotte-minute, delco, dictaphone, durite, frigidaire, gramophone, kalachnikov, klaxon, linotype, lumitype, minerve, minitel, monotype, moulinette, parabellum, photomaton, pianola, polaroid* ('appareil'), *post-it, rimmel, ronéo, sandow, sanisette, scotch, sonotone, ruflette, taser, technicolor, thermos, thyatron, turbulette, typon, velux, zip*, parmi lesquels figurent les sous-groupes des moyens de transport: *deltaplane, mobylette, pédalo, solex, vespa, zodiac*, des produits alimentaires: *bénédictine, esquimau, sucrète*, et des jeux *meccano, scrabble, yoyo / yo-yo*;
- (b) une trentaine de mots désignant des matériaux et des matières, comme des résines, des alliages et des tissus: *box-calf, cellophane, crésyl, duralumin, fibrociment, invar, liberty, mélusine, ozalid, placoplâtre, plexiglas, polaroid* ('feuille'), *pyrex, ripolin, semtex, silicagel, skai, stretch, tartan, téflon, triplex, velcro*, avec les sous-groupes des substances pharmaceutiques: *goménol, mercurochrome, penthotal, thalidomide, vaseline*, des fibres synthétiques: *dacron, nylon, rhovyl, tergal* et des matières plastiques: *bakélite, celluloid, galalithe, rhodoïd*.

De plus, il y a quelques mots que nous n'avons pas pu classer, comme *mûroise* 'fruit issue de l'hybridation naturelle de la ronce et du framboisier' et *solderie* 'magasin où l'on revend à bas prix des invendus qui ont été achetés à moindre coût'. Certes, on reprochera à cette classification d'être peu précise ou incomplète, mais elle permet de distinguer les catégories dominantes. Force est de préciser qu'il existe une classification internationale dite de Nice, adoptée en 1957 en vue de classer les noms de marques, qui définit 34 classes de produits et 11 classes de services d'une manière détaillée qui dépasserait les fins de notre article. À titre d'exemple, citons sa Classe 1 qui inclurait certains mots de notre échantillon:

Classe 1: Produits chimiques destinés à l'industrie et aux sciences; produits chimiques destinés à la photographie; produits chimiques destinés à l'agriculture; produits chimiques destinés à l'horticulture; produits chimiques destinés à la sylviculture; matières plastiques à l'état brut; engrais; compositions extinctrices; préparations pour la trempe de métaux; préparations pour la soudure des métaux; matières pour le tannage de cuirs et peaux d'animaux; adhésifs (matières collantes) destinés à l'industrie; réactifs chimiques autres qu'à usage médical ou vétérinaire; décolorants à usage industriel (INPI 2024c).

Certains noms de marques deviennent des termes. Les noms propres se terminologisent par économie de langue, comme le remarquent Poštolková, Roudný

et Tejnor (1983: 57): par exemple à partir du nom du savant italien Girolamo Cardano, on a créé le syntagme *joint de Cardan* qui est raccourci en *cardan*. Kocourek (1982: 150) mentionne les métonymies fréquentes basées sur les relations entre l'inventeur et l'invention (*rayons roentgen*) et le lieu et le produit (*emmental*). Ainsi, une trentaine d'entrées de notre échantillon contient l'indication du domaine, parmi lesquels figurent, le plus souvent, la technique (*invar*, *fibrociment*, *téflon*, etc.) et l'imprimerie (*linotype*, *lumitype*, *typon*, etc.). D'autres domaines apparaissent moins souvent: le textile (*dacron*, *rhovyl*, *skai*, etc.), la chimie (*crésyl*, *silicagel*), la pharmacie (*mercurochrome*, *penthotal*, *thalidomide*), la couture (*ruflette*), l'électricité (*thyatron*), la mécanique (*delco*), la métallurgie (*duralumin*) et les sports (*tartan*). Cependant, l'indication du domaine n'est pas systématique. Comparons, par exemple, quatre mots désignant des matières synthétiques, dont seulement deux (*dacron*, *rhovyl*) contiennent l'indication du domaine:

DACRON nom masculin. TEXTILE. Fibre synthétique constituée de polyester.

NYLON nom masculin. Matière synthétique à base de résine polyamide.

RHOVYL nom masculin. TEXTILE. Fibre synthétique, résistante et ininflammable, obtenue par filage du chlorure de polyvinyle.

RHODOÏD nom masculin. Matière plastique transparente et quasi incombustible, fabriquée à partir d'acétate de cellulose.

La lexicalisation des noms de marques est rare dans la terminologie officielle. En France, les néologismes terminologiques recommandés par la *Commission d'enrichissement de la langue française*, publiés auparavant au *Journal officiel* et rendus obligatoires pour les agents des services publics de l'État, sont intégrés dans la base de données *FranceTerme* (Ministère de la Culture s.d.). Nous y observons une tendance à remplacer les noms de marques par des termes génériques. Citons par exemple *papillon* (au *Journal officiel* en 2005) qui se substitue à *post-it* dont l'emploi est déconseillé, *chariot* (au *Journal officiel* en 2000) dont l'usage est imposé au lieu de *caddie* (*cart*) ou *pistolet à impulsion électrique* (au *Journal officiel* en 2015) forgé au lieu de *taser*.

4. Traitement des noms de marques dans d'autres dictionnaires généraux

4.1 Comparaison entre les dictionnaires

Pour la comparaison du traitement des noms de marques dans les dictionnaires généraux français, nous avons choisi des ouvrages actuels d'une taille comparable, à savoir le *Dictionnaire de français Larousse* (ci-après *Larousse*), et le *Dico en ligne Le Robert* (ci-après *Le Robert*)¹⁰, les deux disponibles gratuitement en ligne. Selon leurs propres descriptions, le premier compte 135 000 définitions et le second 145 000 définitions régulièrement actualisées.

DAF	Larousse	Le Robert	DAF	Larousse	Le Robert
<i>bakélite</i>	–	NM	<i>pentotal</i>	–	SNM
<i>bénédictine</i>	NM	NM	<i>photomaton</i>	NM	NM
<i>box-calf</i>	SNM	–	<i>pianola</i>	NM	NM
<i>bureautique</i>	NM	NM	<i>placoplâtre</i>	–	NM
<i>caddie</i>	NM	NM	<i>plexiglas</i>	NM	NM
<i>cataphote</i>	NM	SNM	<i>polaroïd</i>	NM	NM
<i>cellophane</i>	NM	NM	<i>post-it</i>	NM	NM
<i>celluloïd</i>	NM	SNM	<i>pyrex</i>	NM	NM
<i>cinémascope</i>	NM	NM	<i>rhodoïd</i>	NM	NM
<i>cinématographe</i>	SNM	SNM	<i>rhovyl</i>	NM	NM
<i>climatiseur</i>	SNM	SNM	<i>rimmel</i>	NM	NM
<i>cocotte-minute</i>	NM	NM	<i>ripolin</i>	NM	NM
<i>crésyl</i>	NM	NM	<i>ronéo</i>	NM	NM
<i>dacron</i>	NM	NM	<i>ruflette</i>	NM	NM
<i>delco</i>	NM	NM	<i>sadow</i>	NM	NM
<i>deltaplane</i>	SNM	SNM	<i>sanisette</i>	NM	NM
<i>dictaphone</i>	NM	NM	<i>scotch</i>	NM	NM
<i>duralumin</i>	NM	NM	<i>scrabble</i>	NM	NM
<i>durite</i>	NM	NM	<i>sentex</i>	NM	–
<i>esquimau</i>	–	NM	<i>silicagel</i>	SNM	–
<i>fibrociment</i>	NM	NM	<i>skai</i>	NM	NM
<i>frigidaire</i>	NM	NM	<i>solderie</i>	SNM	NM
<i>galalithe</i>	SNM	NM	<i>solex</i>	NM	NM
<i>goménol</i>	–	–	<i>sonotone</i>	NM	NM
<i>gramophone</i>	SNM	NM	<i>stretch</i>	NM	NM
<i>invar</i>	NM	NM	<i>sucrette</i>	NM	NM
<i>kalachnikov</i>	SNM	–	<i>tartan</i>	–	NM
<i>klaxon</i>	NM	NM	<i>taser</i>	NM	NM
<i>liberty</i>	NM	NM	<i>techicolor</i>	NM	NM
<i>linotype</i>	NM	NM	<i>téflon</i>	–	NM
<i>lumitype</i>	SNM	NM	<i>tergal</i>	NM	NM
<i>meccano</i>	NM	–	<i>thalidomide</i>	SNM	SNM
<i>mélusine</i>	SNM	–	<i>thermos</i>	NM	NM
<i>mercurochrome</i>	NM	NM	<i>thyatron</i>	SNM	NM
<i>minerve</i>	SNM	–	<i>triplex</i>	NM	NM
<i>minitel</i>	NM	NM	<i>turbulette</i>	–	NM
<i>mobylette</i>	NM	NM	<i>typon</i>	NM	SNM
<i>monotype</i>	SNM	NM	<i>vaseline</i>	SNM	SNM
<i>moulinette</i>	NM	NM	<i>velcro</i>	NM	NM
<i>mûroise</i>	–	–	<i>velux</i>	NM	NM
<i>nylon</i>	NM	NM	<i>vespa</i>	NM	NM
<i>ozalid</i>	NM	NM	<i>yoyo / yo-yo</i>	NM	NM
<i>parabellum</i>	SNM	SNM	<i>zip</i>	NM	NM
<i>pédalo</i>	NM	NM	<i>zodiac</i>	–	NM

Tableau 1: Comparaison de la présence des noms de marques dans les dictionnaires français (NM: nom de marque est indiqué dans l'étymologie; SNM: l'étymologie ne mentionne pas le nom de marque; –: la nomenclature ne contient pas le mot)

Le *Larousse* contient 78 sur 88 mots de l'échantillon. Pour 17 mots, il ne contient aucune référence au nom de marque dans leur étymologie. À part ces 17 entrées, les autres sont systématiquement orthographiées avec la majuscule, y compris dans les exemples. Une graphie spécifique est utilisée pour *CinémaScope*, tandis que le *DAF* et *Le Robert* ne connaissent que la forme *cinémascope*. *Polaroid* y est orthographié sans signe diacritique et l'orthographe anglaise est appliquée pour *Rufflette*, à la différence du *DAF* et du *Robert* qui utilisent *rufflette*. Le mot *esquimau* y réfère seulement aux Inuits et à leur langue.

Seuls 9 mots de notre échantillon sont absents du *Robert*. Dans 10 cas¹¹, l'origine n'est pas indiquée dans le nom de marque. Le mot *minerve* y figure seulement dans le sens appellatif 'appareil orthopédique'. Tous sont écrits avec une minuscule et seule l'étymologie avertit le lecteur qu'il a affaire à une marque déposée. Le mot désignant 'matériau composé d'une plaque de plâtre coulée entre deux épaisseurs de carton' est orthographié *placoplatre* dans *Le Robert*, tandis que la forme *placoplâtre* figure dans le *DAF* et dans le *Larousse*. L'INPI enregistre les marques verbales actives *Placoplatre* et *Placo* dont le déposant est la société anonyme *Placoplâtre*, ce qui complique davantage le traitement lexicographique du mot.

Les trois dictionnaires enregistrent un certain nombre de formes dérivées de ces noms de marques (p. ex. les verbes *klaxonner*, *ronéoter*, *ronéotyper* et *scotcher* dans le *DAF*, le *Larousse* et *Le Robert*, l'adjectif *scotché/e* dans le *DAF* et le *Larousse*, *scrabbleur* dans le *Larousse* et *Le Robert*). Le *Larousse* et *Le Robert* répertorient également un certain nombre d'apocopes des noms de marques originelles (*kalach* et *mob*).

4.2 Le dilemme des lexicographes

Selon l'Organisation mondiale de la propriété intellectuelle (2021), il y aurait 64,4 millions de marques enregistrées dans le monde en 2020 et 13,4 millions de demandes d'enregistrement de marques ont été déposées dans le monde la même année. Bien qu'une partie infime de tous ces noms se lexicalisent, on voit bien le dilemme qui se présente aux lexicographes.

Certains auteurs se sont penchés sur le problème du traitement des noms propres dans les dictionnaires de noms propres (Lecomte-Hilmy 1989), d'autres se sont interrogés sur l'intégration des noms propres dans les dictionnaires de langue (Vaxelaire 2005: par. 3-20), et d'autres encore ont traité directement des choix lexicographiques en matière de noms de marques (Kolaříková et Mudrochová 2019). Selon Fèvre-Pernet et Roché (2005: par. 52-59), sont enregistrés dans les dictionnaires de langue les noms de marques (a) qui sont utilisés comme appellatifs sans que les locuteurs aient une autre dénomination alternative, (b) les noms de marques qui sont employés dans un sens élargi aux objets de la même classe et (c) les noms de marques très notoires et très fréquents dans les corpus utilisés par les lexicographes. On voit bien que ces critères sont assez généraux et que les lexicographes doivent examiner de nombreux cas de manière tout à fait individuelle.

Les 88 items enregistrés dans le *DAF* ne sont en effet nullement une liste exhaustive des noms de marques lexicalisés. Le Tableau 1 démontre l'usage hésitant dans les dictionnaires généraux français, mais ne démontre pas l'absence, dans le *DAF*, de certains mots qui sont classés dans les autres dictionnaires. Ainsi, le *Larousse* et *Le Robert* en ligne enregistrent, par exemple, *abribus*, *fréon*, *frisbee*, *jacuzzi*, *kevlar*, *kleenex*, *néoprène* ou *silentbloc* (tous avec la mention du nom de marque), que l'usager ne trouvera pas dans le *DAF*.

Le lecteur ne trouvera ni dans le *DAF*, et ni dans le *Larousse* ou *Le Robert* quelques mots d'une certaine notoriété, comme *albal*, *eternit*, *lavomatique*. Cependant, il trouvera *audimat*, *botox* et *sopalin* dans *Le Robert*, pour ne citer que quelques exemples. En revanche, le *DAF* contient des mots issus des noms de marques actives ou expirées, sans mentionner cette origine. Par exemple *bateau-mouche* est une marque enregistrée et toujours renouvelée par la Compagnie des bateaux mouches, SA, et *linoléum* est peut-être la première marque lexicalisée, à peine quelques années après l'introduction de ce revêtement de sol. En 1878, Frederick Walton a perdu un procès contre un autre fabricant, pendant lequel le tribunal a déclaré que le mot était tant utilisé qu'il était devenu un terme générique (Pierce s.d.). Sont également incluses *aspirine* et *héroïne*, de vieilles marques enregistrées par Bayer à la fin du XIX^e siècle qui ont été confisquées avec les biens de la compagnie et dépourvues de la protection après la Première Guerre mondiale (Pierce s.d.).

5. Conclusion

Pour certains, l'entrée d'un nom de marque dans un dictionnaire général ou même dans le langage courant pourrait être vue comme «un graal absolu», mais ce passage entraîne aussi une perte de «maîtrise de son image», remarque Lherbier (2022). Cependant, cette entrée est accompagnée de choix délicats qui s'imposent aux lexicographes. Les dilemmes concernent les critères de sélection des lemmes à enregistrer, leur présentation dans le dictionnaire et les informations fournies dans les articles. Ils portent également sur leur orthographe, ainsi que sur celle des mots lexicalisés lorsque ceux-ci diffèrent de la forme officielle des noms déposés. Ils touchent aussi les aspects juridiques de leur inclusion dans le dictionnaire.

L'analyse de notre échantillon, composé de 88 entrées intégrées dans la 9^e édition du *DAF*, a révélé un groupe assez disparate de mots. Ces derniers désignent des appareils, des machines, des produits, des procédés, des outils, des matériaux, des matières et des substances. Formellement, ils sont créés à l'aide d'un large éventail de procédés morphosyntaxiques. Une quinzaine d'entre eux sont des emprunts, parmi lesquels dominent les emprunts à l'anglais. Une analyse rapide portant sur la présence des mots issus des noms de marques dans les dictionnaires français a mis en évidence leur traitement lexicographique assez variable.

En guise de conclusion, évoquons, avec Lherbier (2022), le cas instructif du

mot *Karcher* utilisé en 2005 par Nicolas Sarkozy en qualité de ministre de l'Intérieur, lors de sa visite de la cité de La Courneuve après la mort d'un enfant tué par une balle perdue. La phrase *On va nettoyer au Karcher la cité* a vite gagné du terrain dans le discours politique français (ainsi que les dérivés *karchériser*, *karchérisation*). L'entreprise allemande Kärcher, propriétaire de la marque de nettoyeurs haute pression, a publié un communiqué dans les médias français annonçant que «Kärcher ne souhaite pas que son nom et sa marque soient empruntés par le discours ou pour tout autre sujet qui désignerait autre chose que ses produits» (De Rubercy 2020). Ce souhait n'a pas été entièrement exaucé, comme le démontre Tonti (2021). L'usage de dérivés du mot s'est en effet élargi dans le discours contemporain français pour désigner des activités économiques, sociales, juridiques, écologiques et quotidiennes. Pour le moment, seul le *Wiktionnaire* enregistre *karchériser* avec le sens «nettoyer en profondeur». On peut se demander si les autres dictionnaires de langue vont eux-aussi prendre en compte cette évolution.

Notes

1. «[...] whose function is to steer the choices of consumers and investors and whose use has economic objectives».
2. En plus des marques verbales, ils existent des marques figuratives (dessins, logos, combinaisons de mots et de dessins), des marques sonores, des marques tridimensionnelles, des marques holographiques ou hologrammes, etc. (INPI 2024a).
3. *Robert* désigne un 'sein' dans le langage populaire, alors que le nom de marque désignait à la fin du XIX^e siècle un 'biberon'. *Tabloid*, ayant pour sens 'journal à sensation et à grand tirage, généralement imprimé dans un format réduit', a été emprunté à l'anglais, où *tabloid* était le nom d'une marque déposée de médicament.
4. Il existe aussi des verbes dérivés de noms de marques, comme *scotcher* ou *ronéoter*, enregistrés dans le *DAF*.
5. Selon INPI (2024b), le droit de marques a été instauré par l'adoption de la loi de 1857. Le *Bulletin officiel de la propriété industrielle* est publié à partir de 1884 et l'Office national de la propriété industrielle a été créé en 1902.
6. Sauf indication contraire, les étymologies et les définitions dans cet article proviennent du *DAF*. Quelques-unes ont été abrégées.
7. Le mot *Esquimau* est considéré aujourd'hui comme offensant par les peuples inuits dans toutes ses acceptions et le nom de marque a été abandonné par la société Gervais en raison de sa banalisation. Les bâtonnets glacés ont été rebaptisés et portent le nom de *Kim*. À la différence du *Larousse* et du *Robert*, le *DAF* ne mentionne pas la nature péjorative ou vieillie du mot.
8. Au-delà des acceptions enregistrées par le *DAF*, le mot *triplex* a largement essaimé en français. Comme nom propre, *Triplex* s'emploie par exemple pour une famille de fonts avec trois variantes de base ou pour un autre matériau tri-couches utilisé pour protéger les substrats de la corrosion produit par la compagnie Interpon. Au Canada, il désigne aussi une 'maison comprenant trois logements superposées' selon Le Petit Robert de 2012.

9. Seul le *DAF* mentionne le mot *climatiseur* en tant que nom de marque, tandis que les autres dictionnaires en font un nom commun. La base de données de l'INPI contient cependant la mention de la marque verbale expirée en 1993. Ces incohérences soulèvent des questions sur l'autorité lexicographique en matière de noms de marques.
10. Kolaříková et Mudrochová (2019) traitent d'une centaine de «noms déposés» dans le *Petit Robert* de 2017 en vue de comparer ses équivalents (ou traductions) tchèques.
11. Évidemment, d'autres ouvrages du même éditeur peuvent adopter une approche différente. Ainsi, *Le Petit Robert* de 2012 mentionne l'origine de *deltaplane* ou *cataphote* dans le nom de marque.

Références Bibliographiques

Académie française. s.d. *La 9^e édition du Dictionnaire de l'Académie française.*

https://www.dictionnaire-academie.fr/9e_edition.html [2 février 2025]

Altmanova, J. 2013. *Du Nom Déposé au Nom Commun. Néologie et Lexicologie en Discours.* Milano: EDUCatt.

Altmanova, J. et G. Le Tallec (Éds.). 2019. *Lexicalisation de l'Onomastique Commerciale. Créer, Diffuser, Intégrer.* Bruxelles: Peter Lang.

<https://doi.org/10.3726/b16411>

ATILF — CNRS et Université de Lorraine. s.d. *Trésor de la Langue Française Informatisé.*

<http://www.atilf.fr/tlfi> [2 février 2025]

Botton, M. et J.-J. Cegarra. 1994. *Le Nom de Marque. Création et Stratégies de Marque.* Paris: Ediscience International.

Büchi, E. 1991. Contribution à l'Etude des Déonomastiques Galloromans: Index des Éponymes dans le *FEW. Nouvelle Revue d'Onomastique* 17–18: 139-152.

<https://doi.org/10.3406/onoma.1991.1105>

Catach, L. 2023. Quelques considérations et retours d'expérience sur les éditions numériques du *Dictionnaire de l'Académie française* et du *Dictionnaire de l'Académie nationale de médecine. Études de linguistique appliquée* 211: 283-296.

<https://doi.org/10.3917/ela.211.0029>

Chovancová, K., L. Mešková et S. Krafčíková. 2022. *Lexicalisation des Noms de Marques en Français et en Slovaque.* Paris: L'Harmattan.

De Rubercy, J. 2020. La Marque Kärcher Demande (à nouveau) que son Nom Soit Retiré du Discours Politique.

<https://www.radiofrance.fr/franceinter/la-marque-kaercher-demande-a-nouveau-que-son-nom-soit-retire-du-discours-politique-7508657> [2 février 2025].

Douglas Harper. 2024. *Online Etymology Dictionary.*

<https://www.etymonline.com> [2 février 2025]

Fèvre-Pernet, C. et M. Roché. 2005. Quel Traitement Lexicographique de l'Onomastique Commerciale? Pour une Distinction Nom de Marque/Nom de Produit. *CORELA: Cognition, Représentation, Langage* HS-2.

<http://journals.openedition.org/corela/1198>

<https://doi.org/10.4000/corela.1198>

Fontant, M. 1998. Sur le Traitement Lexicographique d'un Procédé Linguistique: L'Antonimase de Nom Propre. *Cahiers de lexicologie* 73(2): 5-41.

- Galey, B.C.** 1997. *De Mémoire de Marques: Dictionnaire de l'Origine des Noms de Marques*. Paris: Tallandier.
- Galisson, R. et J.C. André.** 1998. *Dictionnaire de Noms de Marques Courants. Essai de Lexiculture Ordinaire*. Paris: Didier Érudition.
- Galkowski, A.** 2022. Markonimy Jako Operatory Taksonomiczne w Terminologii (chremat) onomastycznej. *Onomastica* 66: 299-311.
<https://doi.org/10.17651/ONOMAST.66.20>
- Institut national de la propriété industrielle.** 2024a. Les Différents Types de Marque.
<https://www.inpi.fr/comprendre-la-propriete-intellectuelle/la-marque/les-differents-types-de-marque> [2 février 2025]
- Institut national de la propriété industrielle.** 2024b. Les Marques Peuvent-elles Témoigner de l'Histoire?
<https://www.inpi.fr/innovation/tresors/les-marques-peuvent-elles-temoigner-de-l-histoire> [2 février 2025]
- Institut national de la propriété industrielle.** 2024c. Classifications Internationales.
<https://www.inpi.fr/classifications-internationales> [2 février 2025]
- International Council of Onomastic Science.** s.d. Onomastic Terminology.
<https://icosweb.net/publications/onomastic-terminology> [2 février 2025]
- Jakobson, R.** 1960. Linguistics and Poetics. Sebeok, T.A. (Éd.). 1960. *Style in Language*: 350-377. Cambridge, MA: MIT Press.
- Knappová, M.** 1992. K funkčnímu pojetí systému vlastních jmen. *Slovo a slovesnost* 53(3): 211-214.
- Kocourek, R.** 1982. *La Langue Française de la Technique et de la Science*. Wiesbaden: Brandstetter.
- Kolaříková, D. et R. Mudrochová.** 2019. Les Noms Déposés Issus du Petit Robert et leurs Equivalents en Tchèque. Altmanova, J. et G. Le Tallec (Éds.). 2019. *Lexicalisation de l'Onomastique Commerciale. Créer, Diffuser, Intégrer*: 681-697. Bruxelles: Peter Lang.
- Krško, J.** 2022. Logonyms as a Part of the Linguistic Landscape. *Onomastica* 66: 287-298.
<https://doi.org/10.17651/ONOMAST.66.19>
- Lalanne-Gobet, C.** 2009. *Créer un Nom de Marque et un Nom de Domaine*. Paris: Eyrolles.
- Lapierre, A.** 1989. Problématique des éponymes en français contemporain. Kremer, D. (Éd.). 1989. *Actes du XVIII^e Congrès International de Linguistique et de Philologie Romanes. Tome IV*: 588-597. Tübingen: Max Niemeyer.
- Larousse.** s.d. *Dictionnaire de Français*.
<https://www.larousse.fr/dictionnaires/francais> [10 mars 2025]
- La Stella, E.T.** 1984. *Dizionario Storico di Deonomastica. Vocaboli Derivati da Nomi Propri, con le Corrispondenti Forme Francesi, Inglesi, Spagnole e Tedesche*. Firenze: Leo S. Olschki.
- Lecomte-Hilmy, A.** 1989. Du Statut Linguistique des Noms Propres dans Cinq Dictionnaires Français. *Cahiers de lexicologie* 54(1): 7-32.
- Le Robert.** s.d. *Dico en Ligne*.
<https://dictionnaire.lerobert.com> [20 mai 2025]
- Leroy, S.** 2004. *Le Nom Propre en Français*. Paris: Ophrys.
- Lherbier, L.** 2022. Quand un Nom de Marque Devient un Nom Commun.
<https://yuma-brandthinking.com/2022/06/09/quand-un-nom-de-marque-devient-un-nom-commun> [2 février 2025]
- Marouzeau, J.** 1950. *Les Aspects du Français*. Paris: Masson.
- Ministère de la Culture.** s.d. *FranceTerme*.
<https://www.culture.fr/franceterme> [2 février 2025]

- Nyrop, K.** 1913. *Grammaire Historique de la Langue Française. Tome IV*. Copenhague: Nordisk Forlag.
- Organisation Mondiale de la Propriété Intellectuelle.** 2021. Rapport sur les Indicateurs Mondiaux Relatifs à la Propriété Intellectuelle.
https://www.wipo.int/pressroom/fr/articles/2021/article_0011.html [2 février 2025]
- Petit, G.** 2000. Un Hybride Sémiotique: Le Nom Déposé. *Linguisticae Investigationes* 23(1): 161-192.
<https://doi.org/10.1075/li.23.1.06pet>
- Petit, G.** 2006. Le Nom de Marque Déposée: Nom Propre, Nom Commun et Terme. *Meta* 51(4): 690-705.
<https://doi.org/10.7202/014335ar>
- Pierce, A.** s.d. A Timeline of Brand Names That Became Everyday Words.
<https://llcattorney.com/small-business-blog/business-info/timeline-of-brand-names-that-became-everyday-words> [2 février 2025]
- Poštolková, B., M. Roudný et A. Tejnor.** 1983. *O české terminologii*. Praha: Academia.
- Pruvost, J.** 2021. *Les Dictionnaires Français. Outils d'une Langue et d'une Culture*. Paris: Ophrys.
- Quemada, B.** 1997. Présentation. Quemada, B. (Éd.). 1997. *Les Préfaces du Dictionnaire de l'Académie française 1694–1992: I-XI*. Paris: Honoré Champion.
- République française.** 2024. *Code de la Propriété Intellectuelle*.
https://www.legifrance.gouv.fr/codes/texte_lc/LEGITEXT000006069414 [2 février 2025]
- Schweickard, W.** 1989. Le Traitement des Formes Déonomastiques dans la Lexicographie Française. Kremer, D. (Éd.). 1989. *Actes du XVIII^e Congrès International de Linguistique et de Philologie Romanes. Tome IV*: 242-253. Tübingen: Max Niemeyer.
- Sjöblom, P.** 2016. Commercial Names. Hough, C. (Éd.). *The Oxford Handbook of Names and Naming*. 453-464. Oxford: Oxford University Press.
- Tonti, M.** 2020. *Le Nom de Marque dans le Discours au Quotidien*. Prisme Lexicurel et Linguistique. Paris: L'Harmattan.
- Tonti, M.** 2021. Pour un Dialogue Interculturel: Evolution Sémantique du Nom de Marque Kärcher dans le Discours. *Anales de Filología Francesa* 29: 504-519.
doi:10.6018/analesff.483221
- Vaxelaire, J.-L.** 2005. Nom Propre et Lexicographie Française. *CORELA: Cognition, Représentation, Langage HS-2*.
<http://journals.openedition.org/corela/1239> [07/02/2025]
<https://doi.org/10.4000/corela.1239>
- Vaxelaire, J.-L.** 2019. Lexicographie des Noms de Marques et de Produits. Altmanova, J. et G. Le Tallec (Éds.). 2019. *Lexicalisation de l'Onomastique Commerciale. Créer, Diffuser, Intégrer*: 75-90. Bruxelles: Peter Lang.
- Wooldridge, T.R.** 1998. Projet d'Informatisation du Dictionnaire de l'Académie Française. Quemada, B. et J. Pruvost (Éds.). 1998. *Le Dictionnaire de l'Académie française et la lexicographie institutionnelle européenne*: 309-320. Paris: Honoré Champion.

Aktuelle Aussprachewörterbücher des Deutschen und des Englischen im Vergleich

Krzysztof Nycz, *Institut für Germanistik, Philologische Fakultät,
Universität Rzeszów, Polen*
(knycz@ur.edu.pl)

(<https://orcid.org/0000-0003-2976-1746>)

und

Zygmunt Tęcza, *Institut für Germanistik, Philologische Fakultät,
Universität Rzeszów, Polen*
(ztecza@ur.edu.pl)

(<https://orcid.org/0000-0002-9543-4674>)

Zusammenfassung: In dem vorliegenden Artikel werden insgesamt fünf, zwei deutsche und drei englische Aussprachewörterbücher, die den aktuellen Entwicklungsstand der Phonolexikographie des Deutschen und des Englischen repräsentieren, einer kritisch-vergleichenden Betrachtung unterzogen. Das Ziel dieser Gegenüberstellung besteht nicht nur darin, die Stärken und die Schwächen der einzelnen Werke vor dem Hintergrund der aktuellen lexikographischen Praxis zu erörtern und ggf. auch Impulse für zukünftige Entwicklungen in der Darstellung und Vermittlung von Ausspracheinformationen zu geben, sondern auch, einen interlingualen Diskurs anzuregen, der zur Reflexion über universelle und sprachspezifische Herausforderungen in der phonetischen Standardisierung und Kodifizierung beitragen könnte.

Den Ausgangspunkt bildet eine allgemeine Charakteristik der zu untersuchenden Wörterbücher, bei der Aspekte wie ihre Entstehungsgeschichte, ihre Zielgruppen sowie Konstruktion und Umfang des jeweiligen Wörterverzeichnisses und des einführenden Textteils zur Sprache kommen. Die eigentliche Analyse fokussiert sich auf vier Schwerpunkte: die Auffassung dessen, was als Standardaussprache zu verstehen und zu kodifizieren ist, den Umgang mit phonostilistischen und regionalen Aussprachevarianten, ferner die grafische Gestaltung der verglichenen Nachschlagewerke und schließlich ihre multimedialen Erweiterungen. Im letzten Abschnitt werden die in den genannten Bereichen festgestellten Gemeinsamkeiten und vor allem Unterschiede zwischen den analysierten Wörterbüchern zusammenfassend geordnet und abschließend wird noch kurz ein subjektiver Ausblick auf die Zukunftsperspektiven der Aussprachewörterbücher gewagt.

Schlüsselwörter: AUSSPRACHEWÖRTERBUCH, AUSSPRACHEKODIFIZIERUNG, DEUTSCHE AUSSPRACHE, ENGLISCHE AUSSPRACHE, STANDARDAUSSPRACHE, PHONETISCHER STANDARD, AUSSPRACHEVARIANTEN, PHONOLEXIKOGRAPHIE, DIGITALE LEXIKOGRAPHIE, ELEKTRONISCHE LEXIKOGRAPHIE

Abstract: **Current Pronunciation Dictionaries of German and English Compared.** The following paper takes a critical and comparative look at a total of five dictionaries of

pronunciation, two German and three English, which represent the current state of development in German and English phono-lexicography. The aim of this comparison is not only to discuss the advantages and disadvantages of the individual works in the context of current lexicographical practice and, possibly, to provide ideas for future developments in presenting and conveying pronunciation information, but also to stimulate an interlingual discourse that could contribute to reflection on universal and language-specific challenges in phonetic standardisation and codification.

As a starting point, a general description of the dictionaries under investigation is provided, covering aspects such as their origins, target groups, as well as the construction and extent of the respective word lists and front matters. The actual analysis focuses on four main areas: the concept of what is to be considered and codified as standard pronunciation, the treatment of phono-stylistic and regional pronunciation variants, the graphic design of the dictionaries compared, and finally their multimedia extensions. The last section summarises the similarities and, above all, the differences between the dictionaries analysed in the above areas, and concludes with a brief subjective outlook on the future prospects of pronunciation dictionaries.

Keywords: PRONUNCIATION DICTIONARY, PRONUNCIATION CODIFICATION, GERMAN PRONUNCIATION, ENGLISH PRONUNCIATION, STANDARD PRONUNCIATION, PHONETIC STANDARD, PRONUNCIATION VARIANTS, PHONO-LEXICOGRAPHY, DIGITAL LEXICOGRAPHY, ELECTRONIC LEXICOGRAPHY

1. Einleitung

Im Zeitalter globalisierter Kommunikation und zunehmender Medialisierung stellen Aussprachewörterbücher nach wie vor ein zentrales Instrument zur Normierung und Standardisierung der Aussprache dar. Der vorliegende Beitrag widmet sich einer kritisch-vergleichenden Analyse gegenwärtig in Gebrauch befindlicher orthoepischer Nachschlagewerke des Deutschen und des Englischen. Hierbei werden zwei aktuellste Werke der deutschen Phonolexikographie, nämlich *Deutsches Aussprachewörterbuch* (2009) sowie *DUDEN — Das Aussprachewörterbuch* (Bd. 6 der DUDEN-Reihe, 2023), den drei führenden englischsprachigen Referenzwerken dieser Art in der jeweils jüngsten Ausgabe gegenübergestellt, und zwar dem *CAMBRIDGE English Pronouncing Dictionary* von Daniel Jones (2022), dem *LONGMAN Pronunciation Dictionary* von John Wells (2008) sowie dem *OXFORD Dictionary of Pronunciation for Current English* (2001/2003).

Vier wesentliche Aspekte stehen im Mittelpunkt der Untersuchung: Zum einen wird die jeweilige Auffassung des phonetischen Standards dargestellt, d. h. die Auffassung darüber, was unter Standardaussprache zu verstehen ist. Zum anderen wird der Frage nachgegangen, inwieweit der Unterschied zwischen normativen Vorgaben und sprachlicher Realität im Sinne phonostilistischer bzw. phonopragmatischer Abweichungen von der Standardaussprache, respektive nationale bzw. regionale Aussprachevarianten, in den fraglichen Wörterbüchern berücksichtigt werden. Ein weiterer Schwerpunkt liegt auf der grafischen Gestaltung, die nicht nur als visuelle Repräsentation phonetisch-phonologischer Informationen fungiert, sondern auch deren Zugänglichkeit und Verständlich-

keit beeinflusst. Schließlich werden die multimedialen Komponenten der untersuchten Nachschlagewerke, etwa in Form digitaler Anwendungen und interaktiver Elemente, kritisch hinterfragt — ein Aspekt, der gerade im digitalen Zeitalter sehr stark an Bedeutung gewinnt.

Das Ziel der vergleichenden Betrachtung ist, die Stärken und die Schwächen der einzelnen Werke im Kontext der aktuellen lexikographischen Praxis zu erörtern und Impulse für zukünftige Entwicklungen in der Darstellung und Vermittlung von Ausspracheinformationen zu geben. Durch die Gegenüberstellung deutscher und englischer Ansätze wird zudem ein interlingualer Diskurs angeregt, der zur Reflexion über universelle und sprachspezifische Herausforderungen in der phonetischen Standardisierung und Kodifizierung beiträgt.

2. Gegenstand der Betrachtung. Genese, Ziel und Konstruktion der untersuchten Wörterbücher

Gegenstand des Vergleichs sind im vorliegenden Beitrag — wie bereits erwähnt — insgesamt fünf (zwei deutsche und drei englische) Aussprachewörterbücher, die zum Teil unterschiedliche Traditionen und Ansätze repräsentieren, jeweils aber den aktuellsten Stand der wissenschaftlichen und phonodidaktischen Auseinandersetzung mit der Standardlautung des Deutschen und des Englischen widerspiegeln. Die zu untersuchenden deutschen Aussprachewörterbücher sind:

- *Deutsches Aussprachewörterbuch* von Eva-Maria Krech, Eberhard Stock, Ursula Hirschfeld und Lutz Ch. Anders, erschienen im Jahr 2009 (nachstehend DAWB), und
- *DUDEN — Das Aussprachewörterbuch* (Bd. 6 der DUDEN-Reihe), bearbeitet von Stefan Kleiner und Ralf Knöbl, in seiner jüngsten, aktualisierten 8. Ausgabe von 2023 (nachstehend DU-23).

Die Vorgeschichte des einen wie des anderen Wörterbuchs ist mittlerweile recht lang, sie reicht in beiden Fällen bis in die 1960er Jahre zurück. Das DAWB knüpft an das *Große Wörterbuch der deutschen Aussprache* (GWDA) von 1982 an, das wiederum eine erweiterte Fassung des 1964 in der damaligen DDR zum ersten Mal herausgebrachten *Wörterbuchs der deutschen Aussprache* (WDA) darstellt. Wie im Einleitungsteil des DAWB ausdrücklich betont wird, verfolgt dieses Nachschlagewerk ähnliche Ziele wie das GWDA und es basiert auf denselben konzeptionellen wie methodologischen Ansätzen (vgl. DAWB: 15). Die Kontinuität des DUDEN-Aussprachewörterbuchs nicht nur zu seinem direkten Vorgänger, der ebenfalls von St. Kleiner und R. Knöbl stammenden siebten Auflage aus dem Jahr 2015, sondern auch zu allen früheren Ausgaben (in umgekehrt chronologischer Reihenfolge: von 2005, 2003, 2000, 1990, 1974 und 1962), bei denen Max Mangold die Federführung hatte, ist hingegen allein schon durch die redaktionelle Betreuung und den festen Platz des Wörterbuchs in der zwölfbändigen *DUDEN-Reihe* gewährleistet.

Bei den Aussprachewörterbüchern des Englischen handelt es sich um die folgenden Werke:

- *CAMBRIDGE English Pronouncing Dictionary* von Daniel Jones in seiner 18. Ausgabe von 2022 (nachstehend EPD),
- *LONGMAN Pronunciation Dictionary* von John Wells in seiner 3. Ausgabe von 2008 (nachstehend LPD) und
- *OXFORD Dictionary of Pronunciation for Current English* von Clive Upton, William A. Kretzschmar und Rafal Konopka, das zuerst als Hardcover-Ausgabe im Jahr 2001 und später als Paperback-Version im Jahr 2003 erschien (nachstehend ODP).

Vor allem das erstgenannte Nachschlagewerk blickt auf eine lange Tradition zurück: Bereits 1917 war nämlich die erste Ausgabe von „The Jones EPD“ erschienen, während *LONGMAN Pronunciation Dictionary* 1990 erstmals herausgebracht wurde und *OXFORD Dictionary of Pronunciation for Current English* noch gut ein Jahrzehnt jünger ist. Nach der Erstausgabe Anfang der 2000er Jahre kam die zweite, aktualisierte Ausgabe 2017 unter dem Titel *The Routledge Dictionary of Pronunciation for Current English* (nachstehend RDP) mit Clive Upton und William A. Kretzschmar als Autorenteam auf den Markt.¹

Alle fünf Wörterbücher — d. h. sowohl die deutschen als auch die englischen — definieren ihren Empfängerkreis möglichst weit und nennen dabei explizit zwei verschiedene Zielgruppen, an die sie sich richten: die Mutter- oder ErstsprachlerInnen einerseits und die NichtmuttersprachlerInnen, v. a. Sprachlernende, andererseits (vgl. DAWB: 4. Umschlagseite, DU-23: 11 f., EPD: iii, LPD: xvii, ODP: ix). Die Ersteren, darunter oft Personen, die von Berufs wegen an phonetischer Korrektheit interessiert sind,² suchen meist Auskunft zur Aussprache von Fremdwörtern oder (fremdsprachigen) Eigennamen, ggf. auch zu möglichen Aussprachevarianten einheimischer Wörter; die Letzteren, für die die Sprache des jeweiligen Wörterbuchs eine Fremdsprache ist, benötigen dagegen vor allem Informationen zur Standardaussprache des darin kodifizierten heimischen (Kern-)Wortschatzes, seltener zu dessen Aussprachevarianten (vgl. DU-23: 11 f., LPD: xvii, ODP: ix).

Das Größenverhältnis zwischen den beiden Zielgruppen dürfte sich allerdings für die deutschen und die englischen Wörterbücher unterschiedlich gestalten. Zieht man nämlich zum einen in Betracht, dass sich im Englischen die Aussprache von Wörtern in einem viel geringeren Maße aus deren Schreibung folgern lässt, als es im Deutschen der Fall ist, und zum anderen, dass die Anzahl von Menschen, die Englisch als Fremd- oder Zweitsprache verwenden, im Verhältnis zu der Anzahl der Muttersprachler sehr hoch ist und diese (anders als beim Deutschen) auch deutlich übersteigt, so liegt die Schlussfolgerung nahe, dass grundsätzlich nach einem deutschen Aussprachewörterbuch proportional weitaus weniger NichtmuttersprachlerInnen greifen als nach einem englischen.

Ein Vergleich der Aussprachewörterbücher zeigt weiterhin, dass für das Deutsche und für das Englische z. T. unterschiedliche redaktionelle Ansätze ver-

folgt wurden, die sich allein schon im Umfang der Werke widerspiegeln. Bei den deutschen Wörterbüchern bewegt sich die Gesamtzahl von Seiten, Wort-einträgen und angeführten Transkriptionen in einem recht engen Rahmen. So umfasst das DAWB 1.076 Seiten, etwa 130.000 Lemmata und insgesamt 150.000 Transkriptionen, während DU-23 mit seinen 944 Seiten rund 136.000 Stichwörter und 144.000 Transkriptionen auflistet.

Bei den Wörterbüchern des Englischen ist das Bild differenzierter. Das EPD umfasst etwa 606 Seiten und enthält 80.000 Einträge mit insgesamt über 230.000 Transkriptionen. Im Vergleich dazu zeigt das LPD bei rund 920 Seiten eine deutlich höhere Anzahl von Stichwörtern (138.000) und etwas geringere Gesamtzahl an Transkriptionen, die sich auf ca. 225.000 beläuft. Auffällig ist jedoch vor allem das ODP bzw. das RDP mit entsprechend 1208 und fast 1600 Seiten. Bei ca. 100.000 Einträgen (und zusätzlich 2.000 Neueinträgen im Fall des RDP) werden hier ca. 300.000 Transkriptionen aufgeführt. Der signifikant größere Umfang der beiden Ausgaben im Vergleich zu EPD und LPD resultiert vor allem aus der konsequenten Aufnahme von zwei separaten Transkriptionen für das *British English* und das *American English* — auch dann, wenn beide Varianten identisch sind.

Die Wortlisten der Wörterbücher, bei denen es sich nicht um die Erstausgabe handelt, also: DU-23, EPD und LPD, basieren naturgemäß auf den Wörterverzeichnissen der jeweiligen Vorgängerauflagen. In gewissem Maße betrifft dies auch das DAWB, das sich, wie oben erwähnt, in der Nachfolge des gut ein Vierteljahrhundert älteren GWDA sieht. Die Quellen des in die aktuellen Ausgaben neu aufgenommenen Wortmaterials sind in den deutschen Aussprachewörterbüchern direkter und auch übersichtlicher als in den englischen aufgelistet. DU-23 nennt in diesem Zusammenhang das Rechtschreib- und das Fremdwörterbuch aus dem Hause Duden (vgl. DU-23: 16), in DAWB ist die Quellenliste, da die Aktualisierung und vor allem Erweiterung des Wörterverzeichnisses in diesem Fall ein unvergleichlich größeres Ausmaß haben musste, wesentlich länger. Außer den beiden schon genannten Titeln umfasst sie eine Online-Datenbank der Universität Leipzig, ferner einige Universalwörterbücher und Wörterbücher für Deutsch als Fremdsprache, ein Neologismen- und ein Herkunftswörterbuch, ein rückläufiges Wörterbuch und noch einiges mehr (vgl. DAWB: 17 f.). Demgegenüber fällt die einschlägige Information des in der ersten Ausgabe erschienenen englischen ODP: „[T]he headword list is tied closely to the lists of all the major dictionaries of Oxford University Press“ (ODP: vii) eher spärlich aus.

Die Transkription der Einträge im Wörterverzeichnis reflektiert prinzipiell die Aussprachevariante(n), die das jeweilige Wörterbuch seinen BenutzerInnen als standardgemäß empfiehlt (mehr dazu in Abschn. 3). Bei der Beurteilung und Auswahl der jeweils zu empfehlenden Aussprache werden von den WörterbuchautorInnen oft empirische Quellen analysiert, zu denen diverse Datenbanken, Aufnahmen von Sendungen der elektronischen Medien, im Internet verfügbare Audio- und Videoaufnahmen und nicht zuletzt selbst durchgeführte Akzeptanzuntersuchungen zu konkreten Aussprachevarianten gehören (zu den Letzteren s. Abschn. 4), wobei die Quellen auch in diesem Fall in den deutschen

Wörterbüchern meist expliziter genannt werden als in den englischen (vgl. DAWB: 16, DU-23: 17, LPD: xviii). Wie es dem auch sei, lässt sich wohl das folgende Bekenntnis des EPD auch auf die übrigen Wörterbücher beziehen:

Ultimately, however, the decisions about which pronunciation to recommend, which pronunciations have dropped out of use, and so on, have been based on the editors' intuitions as professional phoneticians and observers of the pronunciation [...] over many years. The opinion of many colleagues and acquaintances has also been a valuable source of advice. (EPD: vi)

Allen Wörterbüchern ist gemeinsam, dass der jeweiligen Wortliste ein mehr oder weniger informativer einführender Textteil (*front matter*) vorausgeht. Im Hinblick auf dessen Beschaffenheit divergieren die untersuchten Nachschlagewerke allerdings erheblich. Bei den Aussprachewörterbüchern des Deutschen fällt vor allem die enorme Ausführlichkeit der Darstellung auf: So umfasst der einführende Teil des DAWB beeindruckende 280 Seiten, was durchaus dem Umfang eines eigenständigen Phonetik-Lehrbuchs entsprechen dürfte. Im Vergleich dazu ist die Einführung im DU-23 mit ca. 150 Seiten zwar kürzer, aber immer noch sehr umfangreich und informativ. Beide Werke setzen damit auf eine intensive theoretisch-phonodidaktische Aufbereitung, die insbesondere für Studierende und Fachleute von hohem Wert ist. In beiden Fällen werden im einführenden Teil, wenn auch in unterschiedlichem Ausmaß und unterschiedlicher Anordnung, u. a. folgende Schwerpunkte behandelt:

- Erklärungen zur Konstruktion des Wörterverzeichnisses bzw. Anleitung zu dessen Gebrauch,
- Kriterien der Aufnahme von Wörtern ins Wörterverzeichnis; Behandlung von Flexionsformen und Wortbildungskonstruktionen,
- die IPA-Transkription und deren jeweilige Handhabung,
- phonologisch-phonetisches Grundwissen: Laut und Phonem, Phonemvarianten; Kriterien der Beschreibung von Vokalen und Konsonanten (distinktive Merkmale); Silbe und Silbengrenze,
- die Standardaussprache, deren historische Entwicklung und aktuelles Verständnis; Variation in der Standardaussprache; im DAWB zusätzlich die Spezifik der (Standard-)Aussprache in Österreich und in der Deutschschweiz, jeweils in einem separaten Hauptkapitel des Einführungsteils erörtert,
- Einzelbeschreibungen der (Standard-)Artikulation sämtlicher Vokale und Konsonanten mit entsprechenden Grafiken des Ansatzrohres,
- Informationen zur Phonem-Graphem-Zuordnung und zur Distribution einzelner Vokale und Konsonanten,
- Aussprachehinweise zu ausgewählten Fremdsprachen.

Somit ermöglichen also die deutschen Aussprachewörterbücher mit ihrem jeweiligen Textteil als einem lehrbuchartigen Werk eine vertiefte Auseinandersetzung mit den Schlüsselfragen der Phonetik und Phonologie.

Im Gegensatz dazu setzen die Aussprachewörterbücher des Englischen auf ein schlankes und pragmatisches Konzept, das vor allem dem schnellen Zugriff auf phonologisch-phonetisches Grundwissen und praktische Transkriptionsregeln dienen soll. Der einführende Textteil ist hier in jedem Wörterbuch vergleichbar lang (bzw. vergleichbar kurz) und umfasst nur ca. 20 Seiten. Folgende Schwerpunkte werden darin behandelt:

- Erklärungen zum Aufbau des Wörterverzeichnisses,
- Informationen zu den Aussprachemodellen bzw. -varianten, die der Kodifizierung zugrunde liegen und im jeweiligen Wörterbuch berücksichtigt werden,
- grundlegende Informationen zur Aussprache von Vokalen, Diphthongen und Konsonanten, zur Wortakzentuierung, zu Assimilationen und Elisionen, Silben und Silbentrennung,
- Hinweise zur Aussprache von fremdsprachigen Lauten,
- die IPA-Transkription und deren jeweilige Handhabung.

Die 18. Ausgabe des EPD enthält zusätzlich eine Sammlung von sechs kurzen Aufsätzen zu unterschiedlichen phonetischen Einzelfragen (zusammengefasst unter der Überschrift *The World of Pronunciation*), die von führenden ExpertInnen im Bereich der Phonetik verfasst wurden. Es handelt sich dabei um folgende Beiträge: *Pronunciation in Spontaneous Speech* von Richard Cauldwell, *Pronunciation for English as a Lingua Franca* von Jennifer Jenkins, *The Daniel Jones Legacy* von Jack Windsor Lewis, *Teaching and Learning Pronunciation* von Jonathan Marks, *The BBC, Its Pronunciation Unit, and 'BBC English'* von Catherine Sangster sowie *Shortcuts in Casual English Pronunciation* von Linda Shockey.

3. Die Auffassung des phonetischen Standards, oder wessen Aussprache kodifiziert wird

In den Aussprachewörterbüchern des Deutschen, DAWB und DU-23, beruht die Auffassung des phonetischen Standards auf einem vielschichtigen Konzept, das einerseits die normativen und regulativen Funktionen der Standardaussprache berücksichtigt, andererseits aber auch ihrer Realitätsnähe und innerer Differenziertheit Rechnung trägt.

In diesem Punkt verfolgen die beiden Wörterbücher zum Teil unterschiedliche Ansätze. So wird in DAWB die Standardaussprache bereits zu Beginn des Einführungsteils als „die mündliche Form der Standardvarietät in der Bundesrepublik Deutschland“ definiert, die „in geografischer und sozialer Hinsicht über eine weite Geltung [verfügt] und [...] insbesondere durch die elektronischen Medien verbreitet [wird]“ (6). Die Standardaussprache sei demnach

- dialektneutral, d. h. sie verzichte bewusst auf mundartliche oder regional gefärbte umgangssprachliche Varianten, sodass sie in allen sozialen Gruppen und von allen MuttersprachlerInnen überregional verstanden werden könne;

- in offiziellen, öffentlichen Situationen bevorzugt zu verwenden;
- durch unterschiedliche Grade der Artikulationspräzision charakterisiert, was bedeute, dass ihre Anwendung auch außerhalb öffentlicher Kontexte möglich sei;
- kodifiziert und erfülle damit eine normative bzw. regulative Funktion; dabei variere der Grad ihrer Verbindlichkeit, die wohlgermerkt in erster Linie für die soziale Gruppe der BerufssprecherInnen gelte (ebd. 7).

Das DU-23 erweitert noch diese Perspektive, indem es zwar den überregionalen Charakter der Standardaussprache unterstreicht, zugleich aber Raum für Varianten zulässt, die im deutschsprachigen Gebiet nur (sub-)national oder großregional verbreitet sind. Trotz ihrer Normativität orientiert sich die so verstandene Standardaussprache eng an der tatsächlichen Sprechwirklichkeit und berücksichtigt neben der als überregional geltenden Aussprache geschulter BerufssprecherInnen auch die in der Bevölkerung üblichen Aussprachevarianten. Somit wird der Standardaussprache auch das Merkmal der inneren Differenziertheit bzw. Variabilität zugeschrieben. Sie zeigt sich darin, dass die in der Bevölkerung tatsächlich verwendete Standardaussprache, selbst in formellen Situationen, in unterschiedlichem Ausmaß von der überregionalen Norm abweicht.

In DAWB werden dagegen, was die innere Differenziertheit bzw. Variabilität angeht, drei konkrete phonostilistische Ebenen der Standardaussprache, die sog. *Grade der Artikulationspräzision* (vgl. 99 ff.), unterschieden:

- (1) hohe bis mittlere Artikulationspräzision (v. a. beim reproduzierenden Sprechen wie Vorlesen von Nachrichten im Radio und Fernsehen);
- (2) sehr hohe Artikulationspräzision (v. a. beim reproduzierenden Sprechen wie z. B. einem feierlichen Vortrag);
- (3) verminderte Artikulationspräzision (v. a. beim freien Sprechen, z. B. bei öffentlich geführten Gesprächen im Radio und Fernsehen).

Als Kodifizierungsgrundlage wird dabei die unter (1) genannte hohe bis mittlere Sprechspannung und Artikulationspräzision gesetzt. Diese Wahl ist übrigens als eine klare Prioritätenverschiebung im Verhältnis zu früheren Aussprachewörterbüchern, etwa dem *Großen Wörterbuch der deutschen Aussprache* (1982), zu bewerten, in dem die sehr hohe Artikulationspräzision die zentrale Bezugsebene war, und findet nun zwangsläufig sowohl bei der Formulierung der Ausspracheregeln im Einführungsteil des Wörterbuchs als auch bei den Ausspracheangaben im Wörterverzeichnis ihren Niederschlag.

In dieser Frage gehen die Autoren des DU-23 teilweise einen anderen Weg. Sie vertreten die Auffassung, dass es für eine möglichst realitätsnahe Dokumentation der deutschen Aussprache unerlässlich sei, den Kreis der als kodifikationsrelevant angesehenen Situationen und SprecherInnen zu erweitern. So seien Sprachformen, die von Millionen deutschen MuttersprachlerInnen in formellen Sprechsituationen — etwa im schulischen Kontext, sowohl von Lehrkräften als auch von SchülerInnen — verwendet und als situationsangemessen eingestuft

werden, als standardsprachlich im Sinne eines ‚Gebrauchsstandards‘ zu werten. Dabei wird direkt auf das im englischsprachigen Raum bereits etablierte Konzept Bezug genommen, den Sprachgebrauch der *educated speakers* als Basis für standard-sprachliche Aussprachewörterbücher und Grammatiken zu nutzen (vgl. DU-23: 30 f.). Diese Vorgehensweise betrachten die Autoren auch für das Deutsche als einen „zeitgemäßen Schritt zu einem erweiterten und damit realitäts- und gebrauchsnäheren Standardsprachkonzept“ (ebd. 31).

Offensichtlich ist ihnen daran gelegen, ein möglichst differenziertes Bild der deutschen Aussprache in formellen Kontexten zu zeichnen. Anders als in DAWB wird hierbei auf eine hierarchische phonostilistische Registeraufspaltung verzichtet. Stattdessen bietet DU-23 in einem speziellen Kapitel einen kompakten und dennoch recht informativen Überblick über relevante Varianten und Variationsphänomene innerhalb der Standardaussprache; alles in allem werden etwa vierzig solcher Erscheinungen auf dem Gebiet der Vokale, Konsonanten und der Realisation von Nebensilben diskutiert. Darüber hinaus sind im Wörterverzeichnis bei einer Reihe von Lemmata auch regional gebräuchliche Aussprachevarianten mit entsprechender Angabe (wie *österr.*, *schweiz.*, *nordd.*, *vorw. nordd.* oder *bes. nordd.*, *südwestd.* etc.) aufgeführt, eine Lösung, die als recht funktional zu bewerten ist und die man in DAWB, das bei einheimischen Wörtern prinzipiell nur eine Transkription angibt, gänzlich vermisst.³

Was nun die Wörterbücher der englischen Aussprache angeht, so fokussieren sie sich traditionell in erster Linie auf das britische Englisch. Dabei galt jahrzehntelang die Aussprache der britischen bzw. südenenglischen Elite der öffentlichen Schulen (*Public School Pronunciation*), später die sog. RP, *Received Pronunciation*, als Modell, auf das sich die orthoepischen Nachschlagewerke bei der Kodifizierung der Aussprache stützten.

In ihrer ursprünglichen Form war die RP eine Art Standard, dem die Aussprache gebildeter SprecherInnen in London und den sog. *Home Counties* zugrunde lag. Ihr Gebrauch war jedoch nicht auf diese Region beschränkt, sondern war im neunzehnten Jahrhundert für die Sprache der Oberschicht im ganzen Land charakteristisch. In England und in Wales gilt die RP weiterhin als Mustermodell für die korrekte Aussprache, insbesondere in der formalen Sprache der gebildeten SprecherInnen (vgl. EPD: v).

Allerdings sind sich die Autoren aller drei aktuellen Aussprachewörterbücher des Englischen einig, dass solch eine konservative Auffassung der RP heutzutage kaum noch haltbar ist. In EPD wird daher programmatisch vorgeschlagen: „The time has come to abandon the archaic name Received Pronunciation“ (ebd.). Stattdessen wird eine modernisierte RP-Variante („a modernized version of RP“), ein breiter angelegtes und leichter zugängliches Aussprachemodell für das britische Englisch vorgeschlagen. In EPD und LPD wird dieses Aussprachemodell als *BBC English* bezeichnet (vgl. EPD: v; LPD: xix).

[T]his is the pronunciation of professional speakers employed by the BBC as news-readers and announcers on BBC1 and BBC2 television, the World Service and BBC Radio 3 and 4, as well as many commercial broadcasting organisations such as ITN. (EPD: v)

Es wird als ein wichtiges Merkmal der modernen RP betont, dass die Sprache der BBC-SprecherInnen von den meisten Menschen nicht mit der hohen sozialen Gesellschaftsschicht und mit Privilegien assoziiert wird, wie dies in der Vergangenheit bei der PSP (*Public School Pronunciation*) und der RP der Fall war (vgl. ebd.). Darüber hinaus wird hervorgehoben, dass die moderne RP überregional ist, d. h. nicht an eine bestimmte Stadt oder Region gebunden.

Das ODP geht diesbezüglich zum Teil seinen eigenen Weg und unterscheidet zwischen *marked* und *mainstream/unmarked RP*. Die *marked RP* ist demnach eine Aussprachevariante, die, wenn sie von den meisten britischen MuttersprachlerInnen gehört wird, dazu führt, dass der Benutzer oder die Benutzerin als altmodisch, affektiert oder präntiös eingestuft wird; sie wird mit einer kleinen Gruppe älterer SprecherInnen der Mittel- und Oberschicht assoziiert, die eine enge Verbindung zum Südosten Englands haben. Währenddessen ist die *mainstream/unmarked RP* eine Aussprachevariante, die mit Ausbildung und Kultiviertheit/Eleganz assoziiert wird, dabei keine besonderen regionalen Untertöne enthält und keine negativen Bewertungen evoziert. Sie wird von den meisten nationalen Radio- und FernsehnachrichtensprecherInnen, von Menschen mittleren Alters und jüngeren Berufstätigen verwendet („broadcast RP“). Die *unmarked RP* ist überregional und nicht an eine bestimmte soziale Schicht gebunden. Es ist die Aussprache eines breiten Spektrums von MuttersprachlerInnen aus vielen Teilen des Landes, die einen sehr unterschiedlichen beruflichen Hintergrund (aber ein eher überdurchschnittliches Bildungsniveau) aufweisen (vgl. ODP: xii). Die *unmarked RP* bildet in ODP die Kodifizierungsgrundlage und entspricht im Allgemeinen der in den zwei anderen Aussprachewörterbüchern beschriebenen modernen RP.⁴

Was nun die amerikanische Aussprache betrifft, so wird sie in allen drei Wörterbüchern mehr oder weniger explizit zu der als *General American* bekannten Sprachvarietät in Verbindung gesetzt, die von der Mehrheit der AmerikanerInnen gesprochen wird, genauer: von denjenigen, die keinen auffälligen östlichen oder südlichen Akzent haben. Das *General American* ist normalerweise bei professionellen SprecherInnen in nationalen Nachrichten- und Informationssendungen zu hören (vgl. LPD: xx; EPD: v–vi).

In EPD und in ODP wird des Weiteren darauf hingewiesen, dass keine einzige Varietät — ob regional oder sozial — als amerikanischer Standard herausgehoben wurde (vgl. EPD: vi; ODP: xiv). So gibt es selbst in den nationalen Medien mit professionell ausgebildeten SprecherInnen auch solche mit regional gefärbter Aussprache. Dennoch könne nach Auffassung des EPD dieses mediale „Network English“ in seiner farblosesten Form schon als eine relativ homogene Variante der gesprochenen Sprache (im Orig. „dialect“) beschrieben werden, welche die fortschreitende Entwicklung der amerikanischen Dialekte widerspiegeln. Darauf Bezug nehmend erklären die EPD-Autoren, in ihrem Wörterverzeichnis die Vielfalt der Aussprachevarianten mit dem geringsten Maß an regionaler oder sozialer Markierung auflisten zu wollen, „while still being sensitive to the traits of the individual word“ (EPD: vi).

Regionale Varianten, und zwar sowohl für das britische als auch für das amerikanische Englisch, werden in keinem der analysierten englischen Wörterbücher aufgeführt. In EPD und in ODP wird darauf explizit hingewiesen, in dem Letzteren mit folgender Begründung:

It would of course be both impracticable and confusing to attempt to present transcriptions for even a fraction of the variety of regional and social accents which characterize British and American English. For both pragmatic and pedagogic reasons model accents have to be sought, transcriptions of which may be generally accepted as embodying the major elements of the pronunciation-types under consideration. (ODP: xix)

Mit praktischen Gründen wird auch von den Autoren des EPD argumentiert, wobei sie eher den enormen Arbeitsaufwand in den Vordergrund stellen, den die Aufnahme von regionalen Aussprachevarianten ins Wörterbuch erfordern würde:

A pronouncing dictionary that systematically presented the pronunciations of a range of regional accents would be very valuable, but it would be very much bigger than the present volume and the job of ensuring an adequate coverage which treated all accents as equally important would have taken many years. (EPD: vi)

Die (Nicht-)Berücksichtigung von regional verbreiteten Spielarten der jeweiligen Aussprache ist als einer der markantesten Unterschiede zwischen den beiden deutschen und den drei englischen Wörterbüchern anzusehen, die hier Gegenstand des Vergleichs sind. Zieht man jedoch in Betracht, dass Deutsch hauptsächlich in drei Ländern mit insgesamt etwa 100 Millionen Einwohnern gesprochen wird, Englisch dagegen außer in Großbritannien und den USA auch in Irland, Kanada, Australien, Neuseeland sowie in einigen weiteren Dutzend Ländern der Welt als Mutter- und/oder Amtssprache in Gebrauch ist, so muss der Anspruch, die gesamte Vielfalt von gegenwärtig vorkommenden Aussprachevarianten des Englischen zu erfassen, in der Tat als recht unrealistisch bewertet werden.

4. Grafische Gestaltung und multimediale Komponenten

4.1 Grafische Gestaltung

Die hier untersuchten Aussprachewörterbücher sowohl des Deutschen als auch des Englischen sind im Vergleich zu den Wörterbüchern früherer Generationen durchaus fortschrittliche Nachschlagewerke, die durch eine klare, benutzerfreundliche und didaktisch orientierte grafische Gestaltung auffallen und neuartige, multimediale Komponenten — dem digitalen Zeitalter entsprechend — enthalten. Bis auf das ODP setzen alle Werke in ihrem jeweiligen Wörterverzeichnis visuell hervorstechende Informationskästen ein, in denen das phonetisch-phonologische

Grundwissen vermittelt und auf die jeweiligen Aussprachekonventionen bzw. -besonderheiten aufmerksam gemacht wird.

In den beiden Aussprachewörterbüchern des Deutschen werden dabei ähnliche gestalterische Lösungen verwendet. Im Wörterverzeichnis des DAWB und des DU-23 finden sich nämlich ca. 190 bzw. knapp 180 farbige Infokästen (vgl. Abb. 1), die jeweils an das direkt vorausgehende Stichwort anknüpfen, auf Aussprachebesonderheiten oder -schwierigkeiten hinweisen und — wie in der Anleitung von DAWB versprochen — „Anregungen zur produktiven Erschließung weiterer Ausspracheformen durch den Nutzer“ (279) liefern.



Abb. 1: Infokästen in DAWB (links) und in DU-23 (rechts)

In Wirklichkeit ist aber die Anzahl der Infokästen in DAWB wesentlich geringer, da viele davon, mitunter mit leichten Anpassungen, mehrfach wiederholt werden. So kommen beispielsweise die beiden Sonderregeln: (1) dass [h] im Silbenauslaut vor dem Schwa-Laut und vor den Suffixen *-en*, *-er*, *-ig*, *-ung* nicht gesprochen wird, und (2) dass das Graphem *e* der Endung *-ie* in den Fremdwörtern, in denen diese Endung als [i:] zu realisieren ist, im Singular zwar nur als Dehnungszeichen fungiert, im Plural aber schon als Schwa-[ə] gesprochen werden muss, insgesamt je ca. zwanzigmal vor. Eine Reihe weiterer Informationen, etwa zu Stimmlosigkeitsassimilation von Konsonanten, Konsonantengemination an Silben- und Wortgrenzen, Frikatisierung von [t] vor einem unsilbischen i-Laut, unsilbischer Realisation bestimmter Vokale in manchen Fremdwörtern, auch zu Auslautverhärtung sowie vokalischer Realisation des Suffixes *-er*, die bei einer Silbengrenzverschiebung (z. B. infolge weiterer Suffigierungen) zur konsonantischen Realisation wechseln kann, wird über zehnmal wiederholt.

Die Informationsmodule des DU-23 wirken zwar auf den ersten Blick ähnlich wie jene des DAWB, unterscheiden sich jedoch von ihnen grundlegend in ihrer Struktur und dem vermittelten Inhalt. Sie werden auch jeweils nur einmal eingesetzt; im Gegensatz zu DAWB gibt es also keine Wiederholungen. Rund 86 von insgesamt 177 Infokästen präsentieren prozentuale Ergebnisse einer Online-Umfrage zur Akzeptanz alternativer Aussprachevarianten ausgewählter Wortformen, bei der Angemessenheitsurteile von 573 Probanden aus verschiedenen Regionen Deutschlands, Österreichs und der deutschsprachigen Schweiz berücksichtigt wurden. Dabei beziehen sich die meisten dieser Statistiken auf die Aussprache von konkreten Fremd- und Lehnwörtern, und gut ein Drittel davon thematisiert Schwankungen des Wortakzents (vgl. Abb. 2).

Bei den wenigen einheimischen Wörtern, deren Aussprachepräferenzen statistisch erhoben wurden, lag der Fokus der Untersuchung vor allem auf der Wahl zwischen einem langen gespannten und einem kurzen ungespannten Vokal (wie bei *Nische*, *Walnuss* und einigen weiteren) bzw. auf der rein qualitativen [ɛ:]/[e:]-Schwankung (wie bei *Mädchen*), und lediglich in einem Fall handelte es sich um den variablen Wortakzent (*liebkoosen*).

Anis nordd./westmd. vorw.:
a'ni:s, österr./ostmd./
schweiz./südd. vorw.: 'a:nis,
auch: 'anis

Anis

Umfrage: Wortakzent

• Norden und Mitte des deutschen Sprachraums

- a'ni:s	+81%	4%	-15%
- 'a:nis	+29%	12%	-59%

• Süden des deutschen Sprachraums

- a'ni:s	+31%	15%	-54%
- 'a:nis	+82%	9%	-9%

Abb. 2: DU-23: Infokästen zum Lemma *Anis*

Die übrigen Infokästen, die keine statistischen Daten enthalten, liefern jeweils einen Kommentar zur Wortaussprache, der aber im Unterschied zu DAWB nur vereinzelt präskriptiven Charakter hat; in den meisten Fällen werden Aussprachevarianten kurz diskutiert und problematisiert, und sehr oft wird auch hier auf konkrete regionale Differenzen hingewiesen.⁵

Was nun die Aussprachewörterbücher des Englischen angeht, so bieten zwei von ihnen, EPD und LPD, ähnliche gestalterische Lösungen. In ihrem jeweiligen Wörterverzeichnis sind nämlich auch farbige Infokästen, die sog. *information panels*, verstreut.

Da die Korrespondenz zwischen Graphie und Aussprache im Englischen sehr oft inkonsistent ist und die Schreibung eines Wortes häufig keinen eindeutigen Hinweis darauf gibt, wie dieses auszusprechen ist, werden sowohl in LPD als auch in EPD zunächst die sog. *Pronouncing-the-letter-* bzw. *Spelling-to-sound*-Informationsmodule verwendet. Sie sind jeweils der Liste von Einträgen vorangestellt, die mit dem betreffenden Buchstaben des Alphabets beginnen. Der Einsatz gerade solcher Module wird von den Autoren des LPD wie folgt begründet:

English spelling is notorious for its shortcomings. Knowing the orthography of a word does not enable one to predict its pronunciation with any confidence. Nevertheless, certain general principles do govern the relationship between spelling and sound (grapheme and phoneme), even though they may be subject to exceptions and uncertainties. Although many handbooks of English pronunciation ignore them entirely, on the implicit grounds that these rules are so chaotic that it is better to learn the pronunciation of each new word separately, it nevertheless seemed helpful for LPD to offer the user something rather than nothing. (LPD: xviii)

Dementsprechend werden zu jedem Graphem bzw. einer Graphemkombination präzise Ausspracheregeln oder klare Hinweise gegeben, wie diese in unterschiedlicher lautlicher Umgebung auszusprechen sind. Dabei wird in jedem Fall auf sämtliche Unregelmäßigkeiten und etwaige Ausnahmen hingewiesen.

Obwohl die fraglichen Infokästen der beiden Wörterbücher ähnlich aussehen und ähnliche Inhalte vermitteln, sind sie teilweise unterschiedlich aufgebaut. Während in LPD die entsprechenden Grapheme und ihre möglichen Kombinationen samt Ausspracheunregelmäßigkeiten bzw. -ausnahmen in einem Modul behandelt werden, sind diese Inhalte in EPD in separaten Infokästen untergebracht. So findet man in EPD direkt unter A das Modul *Pronouncing the letter A* (s. Abb. 3), auf den folgenden Seiten kommen aber noch vier weitere Module zu verschiedenen Graphemkombinationen, nämlich *Pronouncing the letters AE*, *Pronouncing the letters AEO*, *Pronouncing the letters AI*, *AY* und *Pronouncing the letters AU*, *AW*. Dies hat zwei Folgen: Da in LPD sämtliche Ausspracheregeln zu einem Graphem bzw. dessen Kombinationen in nur einem Infokasten zusammengestellt sind, fallen manche dieser Infokästen recht umfangreich aus (siehe z. B. *O Spelling-to-sound* oder *S Spelling-to-sound*). Im Gegensatz dazu haben die Informationsmodule in EPD — da sie bestimmte Einzelfragen in separaten Modulen behandeln — zwar einen geringeren Umfang, dafür sind sie aber teil-

weise informativer: Zum einen werden in (fast) jedem Modul Angaben zur amerikanischen Aussprache gemacht und zum anderen werden in den Abschnitten *In addition* und *In weak syllables* zusätzliche Aussprachevarianten diskutiert. Die Anzahl von Infokästen dieser Art variiert dementsprechend von 26 in LPD bis insgesamt 69 in EPD.

1
abduct

A

A

Pronouncing the letter A

→ See also **AE, AEO, AI/AY, AU/AW**

The vowel letter **a** has two main strong pronunciations linked to spelling: a 'short' pronunciation /æ/ and a 'long' pronunciation /eɪ/. In the 'short' pronunciation, the **a** is usually followed by a consonant which closes the syllable, or a double consonant before another vowel, e.g.:

tap	tæp
tapping	ˈtæp.ɪŋ

The 'long' pronunciation usually means the **a** is followed by a single consonant and then a vowel, e.g.:

tape	teɪp
taping	ˈteɪ.pɪŋ

When there is an **r** in the spelling, the strong pronunciation is one of three possibilities: /ɑː/ (US) /ɑːr/ (UK) /er/ or /æ/ (US) /eɪ/, e.g.:

car	kɑː (US) kɑːr
care	keə (US) keɪ
carry	kæ.rɪ (US) keɪ.rɪ, kær.i

In addition

There are other vowel sounds associated with the letter **a**, e.g.:

ɑː	father /ˈfɑː.ðər/ (US) /ˈfɑː.ðɜː/
ɑː (US) æ	bath /bɑːθ/ (US) /bæθ/
ɪ (US) ɑː	swan /swɒn/ (US) /swɑːn/
ɔː (US) ɑː, ɔː	walk /wɔːk/ (US) /wɑːk/
	warm /wɔːm/ (US) /wɔːrɪm/

And, in rare cases:

e	many /ˈmeni/
---	--------------

In weak syllables

The vowel letter **a** is realized with the vowels /ə/ and /ɪ/ in weak syllables, and may also not be pronounced at all in British English, due to compression, e.g.:

above	əˈbʌv
village	ˈvɪ.lɪdʒ
necessary	ˈnes.ə.səri (US) -ser.i

a indefinite article: strong form: eɪ; weak form: ə
 Note: Weak-form word. The strong

abnegate /ˈæb.nɪ.ɡeɪt/, in which it is stressed, and **abduct** /əbˈdʌkt/, where it is unstressed.

Abbeville in France: ˈæb.vɪl, (US) ˈæbˈvɪl; in US: ˈæb.ɪ.vɪl
abbrev. A- ˈæb.i -s -z

Abb. 3: EPD: Infokästen zu *Pronouncing the letter A*

Das phonetisch-phonologische Wissen wird in LPD in rund 40 sogenannten *Language Panels* vermittelt (vgl. Abb. 4). Die behandelten Themen reichen von der Definierung phonetisch-phonologischer Grundbegriffe wie z. B. Phonem und Allophon, Diphthong, Affrikate, Geminate, über die Beschreibung phonetischer Einzelfragen wie Aspiration, Assimilation, Elision, Glottisschlag, bis hin zu supra-segmentalen Erscheinungen wie Akzentuierung. Einige wenige Module behandeln Schwerpunkte wie australisches Englisch oder die Aussprache von Sonderzeichen, die in E-Mails oder auf Webseiten verwendet werden, wie @, #, /. Der Übersichtlichkeit halber sind alle *Language Panels* in einem Verzeichnis mit dem Titel *Index to notes on pronunciation and phonetics* mit Seitenangabe aufgelistet, das sich im einleitenden Teil des Wörterbuchs befindet.

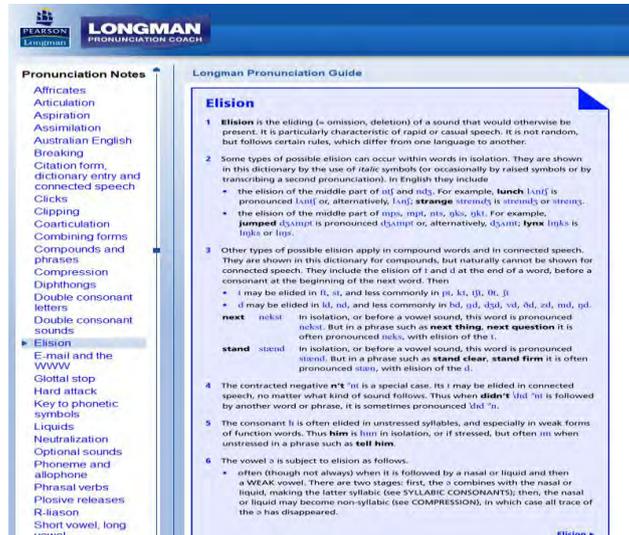


Abb. 4: LPD: *Language Panels* zur phonetisch-phonologischen Terminologie (CD-ROM)

Ähnlich wie in dem deutschen Aussprachewörterbuch DU-23 sind auch in LPD Infokästen mit Statistiken vorhanden (vgl. Abb. 5). Sie bringen prozentuale Ergebnisse von fünf Meinungsumfragen, den sogenannten *opinion polls* (aus den Jahren 1988, 1993, 1998, 1999-2002 und 2007), zu Präferenzen in Bezug auf bestimmte Aussprachevarianten ausgewählter Wörter. Für Einzelheiten zu den Umfragen sowie dazu, welche Wörter in welcher Umfrage enthalten waren, wird auf die Website des Autors verwiesen (vgl. LPD: xviii).

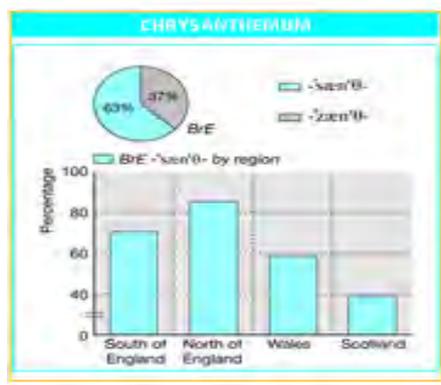


Abb. 5: LPD: *Preference Polls* zur Akzeptanz bestimmter Aussprachevarianten (CD-ROM)

EPD hingegen verzichtet in seiner aktuellen, 18. Auflage auf die Infokästen zu den phonetisch-phonologischen Grundfragen. Stattdessen wird am Ende des Wörterverzeichnisses ein Glossar mit der einschlägigen Terminologie angeboten. Darüber hinaus finden sich in EPD über das gesamte Wörterverzeichnis verstreut Verwendungshinweise, sogenannte *usage notes* (vgl. Abb. 6). In den meisten Fällen werden hier kurz Aussprachevarianten diskutiert oder auch komplexere Ausspracheregeln erläutert (vgl. EPD: vii).

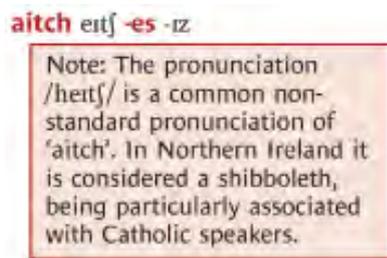


Abb. 6: EPD: Infokasten vom Typ *Usage Note*

Das dritte der behandelten englischen Aussprachewörterbücher: ODP, das, wie bereits in Abschn. 2 erwähnt, als einziges in seinem Wörterverzeichnis für jedes Lemma sowohl britische als auch amerikanische (mitunter gleich lautende) Aussprachevarianten aufführt, geht auch bei der grafischen Gestaltung seinen eigenen Weg. Das Layout des gesamten Wörterbuchs ist so klar und einfach wie möglich gehalten, es gibt darin weder Grafiken noch Informationsmodule, und die laufende Wortliste wird durch keine Sonderelemente unterbrochen.

4.2 Multimediale Komponenten

Mit der Integration der unter 4.1 beschriebenen Informationsmodule in das Wörterverzeichnis der untersuchten Aussprachewörterbücher des Deutschen und des Englischen wurde ein grafisches Mittel eingesetzt, das nicht nur zusätzliche Informationen zur Aussprache vermittelt, sondern auch die visuelle Monotonie einer herkömmlichen Wortliste auflockert. Darüber hinaus bieten aber die Wörterbücher noch einige weitere innovative medientechnische Funktionen, darunter Audiomaterial, was bei diesem Wörterbuchtyp von besonderem Interesse und hoher Relevanz für die BenutzerInnen ist.

So ist DAWB mit einer Audio-DVD ausgestattet, auf der die standardisierte (bundesdeutsche) Aussprache ausgewählter Wortbeispiele aus dem Einführungsteil präsentiert wird. Sie enthält insgesamt 98 Audiodateien im wav-Format. Im Wörterbuch selbst sind die Abschnitte, für die Audiodateien zur Verfügung stehen, deutlich gekennzeichnet und durchnummeriert. Die Entscheidung der AutorInnen, insgesamt ca. 2.000 Worteinheiten und Wortgruppenlexeme aus dem Textteil für die Vertonung auszuwählen, ist wohldurchdacht, denn

gerade diese Beispiele illustrieren in den einführenden Kapiteln verschiedene Ausspracheregeln und besitzen somit einen hohen Verallgemeinerungs- bzw. Analogiewert.

Darüber hinaus ist die gesamte Druckausgabe des DAWB auch als E-Book im PDF-Format verfügbar, sodass es ebenso gut auf einem PC oder anderen elektronischen Geräten genutzt und leicht durchsucht werden kann. In einem engen Zusammenhang mit dem Wörterbuch steht die online zugängliche *Deutsche Aussprachdatenbank* (DAD), die auf dem Wörterverzeichnis des DAWB basiert und dieses durch Ergänzungen, Erweiterungen und Korrekturen weiterführt. Nach Angaben der Projektbetreibenden soll zudem eine (synthetische) Vertonung der in der DAD enthaltenen Stichwörter geplant sein.⁶

Demgegenüber ist das DUDEN-Aussprachewörterbuch nicht nur als Hardcover-Buch und PDF-Datei, sondern auch — wie schon die Vorgängerausgabe von 2015 — als eine richtige Software (Desktop-App) erhältlich, die von der Homepage des Verlages im Rahmen der sog. *Duden-Bibliothek* gegen entsprechendes Entgelt heruntergeladen werden kann.⁷

Die aktuelle, 8. Auflage wurde nach Angaben des Verlags um über 4.000 Stichwörter erweitert und enthält somit 144.000 Aussprachen zu Wörtern und Eigennamen, einschließlich der im Deutschen gebräuchlichen Fremdwörter und fremdsprachigen Namen. Alle Stichwörter wurden vertont und die Audios stehen als Download zur Verfügung, wobei die Vertonungen nur beim Kauf der Printausgabe oder der Software verfügbar sind. Da die neue Auflage erst seit Oktober 2023 auf dem Markt ist, muss noch mit gewissen technischen Unzulänglichkeiten gerechnet werden; darauf wird auch auf der Webseite des Verlages direkt hingewiesen. Einige Audiodateien sind nämlich noch nicht verfügbar, sie sind in der Datenbank mit einem grauen Lautsprechersymbol gekennzeichnet.⁸ Auf jeden Fall ist aber die Bereitstellung von Audios zu allen Einträgen des Wörterbuchs als ein großer Fortschritt im Vergleich zu der Vorgängerauflage zu werten, da jene insgesamt nur 13.000 Vertonungen enthält, was knapp 10 % aller Stichwörter ausmacht.

Von den drei englischen Aussprachewörterbüchern bieten zwei, nämlich EPD und LPD, neben der gedruckten Buchausgabe zusätzlich Audiodateien auf einer CD-/DVD-ROM an.⁹ In beiden Fällen enthält die CD-/DVD-ROM die Vertonung jedes Wörterbucheintrags jeweils in zwei nationalen Aussprachevarianten, das heißt in britischem und amerikanischem Englisch. Beide CD-/DVD-ROMs bieten außerdem interaktive Suchfunktionen, mit denen die Einträge nach verschiedenen Kriterien durchsucht werden können. So kann man z. B. die Wortliste nach bestimmten Lauten oder Lautsequenzen filtern, um gezielt ähnliche Wörter oder Minimalpaare ausfindig zu machen. In LPD erscheinen in der Registerkarte *Longman Pronunciation Guide* wiederholt sämtliche *Spelling-to-sound*-Informationsmodule sowie die Module zur phonetisch-phonologischen Terminologie. Auch die *Preference Polls* werden erneut angeboten, diesmal in digitaler Form.

Was die beiden Werke von den Aussprachewörterbüchern des Deutschen stärker unterscheidet, sind interaktive Übungsmodule verschiedener Art (s. Abb. 7 u. 8), die es u. a. ermöglichen, die eigene Aussprache aufzuzeichnen und sie mit

der zugänglichen modellhaften britischen oder amerikanischen (Standard-)Aus-
sprache zu vergleichen. Dies ist eine technische Lösung, die es in den deutschen
Wörterbüchern nicht gibt. Und was die englischen Aussprachewörterbücher für
ihre BenutzerInnen noch interessanter macht, ist die Möglichkeit, in Hunderten
von interaktiven Übungen die phonetische Lautschrift, die Perzeption und die
Identifikation von Lauten und Lautkombinationen sowie die Laut-Graphem-
Korrespondenz zu trainieren. Bei LPD ist es darüber hinaus möglich, Handouts
mit Übungen aus dem *Teachers' Resource Centre* auszudrucken.

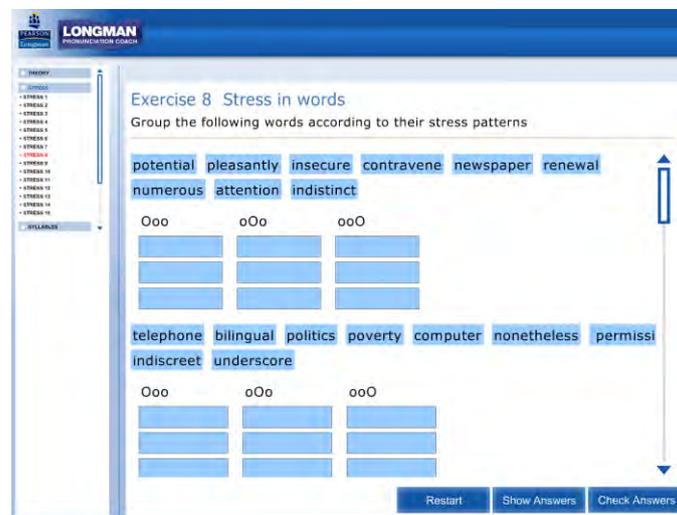


Abb. 7: LPD: Interaktive Übungen (1)

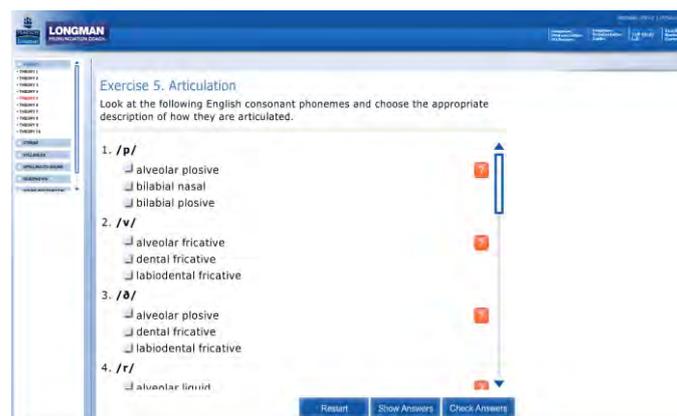


Abb. 8: LPD: Interaktive Übungen (2)

Im Jahr 2015 wurde auch eine digitale Version der 18. Ausgabe des EPD veröffentlicht. Es handelt sich dabei um eine App, die für das iPhone und das iPad von Apple verfügbar ist und über den Apple Store vertrieben wird. Eine Android-Version erschien dann zwei Jahre später.

Das einzige der drei untersuchten Aussprachewörterbücher des Englischen, das kein Begleitmedium in Form einer CD-/DVD-ROM mitliefert, ist das ODP. Andererseits ist es aber auch das einzige, das in seiner zweiten Ausgabe von 2017 (*The Routledge Dictionary of Pronunciation for Current English*, RDP) zugleich als E-Book erhältlich ist.

5. Zusammenfassung und Ausblick

Die untersuchten Aussprachewörterbücher des Deutschen und des Englischen sind ausgereifte Nachschlagewerke, die nicht nur durch ihren Umfang, sondern mehrheitlich auch durch ihre grafische Gestaltung und ihre multimedialen Komponenten überzeugen.

Die festgestellten erheblichen Unterschiede im Umfang des einführenden Textteils (*front matter*) zwischen den Aussprachewörterbüchern des Deutschen und denen des Englischen reflektieren zum Teil unterschiedliche redaktionelle Ansätze. Während die Einführungskapitel der deutschen Wörterbücher — mit insgesamt 150 bis 280 Seiten — als quasi-lehrbuchartige Werke eine vertiefte Auseinandersetzung mit den phonetisch-phonologischen Fragen ermöglichen, liegt bei den Aussprachewörterbüchern des Englischen der Fokus stärker auf der direkten Anwendung und dem schnellen Nachschlagen, während die theoretische Fundierung eine geringere Priorität hat. Auch scheint in den englischen Wörterbüchern weniger Wert darauf gelegt zu werden, den BenutzerInnen genauere erstellungstechnische Hintergrundinformationen, etwa über die Zusammenstellung des Wörterverzeichnisses oder empirische Quellen für die zu transkribierenden Aussprachevarianten, in dem jeweiligen Einführungsteil zu bieten.

Gemeinsam ist dafür allen untersuchten Werken, dass sie — abgesehen von der einzelsprachlich bedingten Spezifik — den (ortho)phonetischen Standard im Vergleich zu den jeweils früheren Ansätzen weniger elitär, also deutlich breiter und damit auch realitätsnäher definieren. Sollte hier dennoch unter diesem Gesichtspunkt hierarchisiert werden, so lässt sich wohl am ehesten das DAWB hervorheben, und zwar als etwas konservativer als die übrigen Wörterbücher, da es die Standardaussprache ausdrücklich in erster Linie mit offiziellen Kommunikationssituationen in Verbindung setzt und den präskriptiven bzw. regulativen Charakter seiner Transkriptionen stärker betont (Abschn. 3).

In unterschiedlichem Ausmaß bzw. mit unterschiedlicher Tiefgründigkeit wird in den fraglichen deutschen und englischen Wörterbüchern der Differenziertheit der jeweiligen Aussprache, darunter vor allem ihrer regionalen Variation Beachtung geschenkt. Dass die Aussprachewörterbücher des Englischen auf die Darstellung von regional oder gar großregional verwendeten Aussprache-

varianten weitgehend verzichten, lässt sich jedoch, wie im Obigen (Abschn. 3) erörtert, durch die außerordentlich große Ausbreitung und Vielfältigkeit der englischen Sprache durchaus plausibel erklären.

Bezüglich der grafischen Gestaltung ist vor allem festzuhalten, dass gleich vier der fünf untersuchten Wörterbücher in ihrem jeweiligen Wörterverzeichnis die sog. Infokästen (*information panels*) mit diversen Inhalten einsetzen, die visuell attraktiv wirken und allesamt nützliche Ergänzungsinformationen bringen. Als Einziges fällt hierbei das ODP (auch in der jüngeren Routledge-Ausgabe, RDP, von 2017) aus dem Rahmen, da es sowohl in seinem Einführungsteil als auch in der Wortliste auf jegliche grafische Mittel verzichtet und überhaupt ein sehr traditionelles Erscheinungsbild bietet, das aus der heutigen Sicht schon recht spartanisch wirkt.

Ein entscheidender Aspekt bei der Erstellung und Analyse von Aussprachewörterbüchern ist heutzutage die fortschreitende technologische Entwicklung, die die Lexikographie zunehmend in digitale Formate überführt. Mit Ausnahme des ODP/RDP, das auch in dieser Hinsicht den übrigen Wörterbüchern nachsteht, enthalten jetzt schon die untersuchten Werke multimediale Erweiterungen, sei es als eine richtige Desktop- oder Smartphone-Anwendung mit dem gesamten Inhalt des Papierwörterbuchs und Audios zu allen Stichwörtern (DU-23, EPD), sei es in Form einer beigefügten CD-/DVD-ROM, die entweder Audiodateien zu ausgewählten Stichwörtern enthält (DAWB), oder aber die Vertonung aller Stichwörter und darüber hinaus erweiterte Suchfunktionen sowie interaktive Ausspracheübungen bietet (EPD, LPD). Hinzu kommt die Fortführung des DAWB durch eine Online-Aussprachedatenbank (vgl. Abschn. 4.2).

Angesichts der rasanten Fortschritte in der Sprach- und Spracherkennungstechnologie sowie der zunehmenden Verfügbarkeit interaktiver Online-Wörterbücher ist es jedoch offensichtlich, dass die derzeit angebotenen Printausgaben mit multimedialen Komponenten (wobei Datenträger wie CD- oder DVD-ROM bereits als überholt gelten müssen) lediglich eine Übergangsphase in der Entwicklung der Phonolexikographie darstellen.¹⁰ Digitale Plattformen bieten nämlich nicht nur den Vorteil der ständigen Aktualisierung, sondern auch die Möglichkeit einer noch anschaulicheren und flexibleren Darstellung phonetisch-phonologischer Informationen, etwa durch interaktive IPA-Transkriptionen oder KI-gestützte Aussprachehilfen.

Vor diesem Hintergrund ist davon auszugehen, dass die Weiterentwicklung von KI-Technologien tiefgreifende Auswirkungen auf die Zukunft von Aussprachewörterbüchern haben wird. KI-gestützte Systeme können eine nahezu unbegrenzte Menge an Sprachdaten auswerten, fortlaufend aktualisieren und auch individualisierte Rückmeldungen zur Aussprache geben, indem sie Benutzeraufnahmen mit Standardvarianten vergleichen. Daraus erwächst die Möglichkeit, dass Aussprachewörterbücher, heute noch primär statische Referenzwerke, sich mit der Zeit zu speziellen interaktiven Informations-, Lern- und Trainingsplattformen weiterentwickeln (mehr zur aktuellen Beschäftigung mit den Auswirkungen der KI auf die Lexikographie siehe z. B. De Schryver 2023: 355 ff., McKean und Fitzgerald 2024: 7 ff., Ptasznik und Lew 2025: 140 ff.).

Denkbar ist aber auch ein anderes Szenario. Zieht man nämlich in Betracht, dass online zugängliche renommierte Universalwörterbücher wie die deutschen *DWDS* und *Duden online*, die britischen *Oxford English Dictionary (OED)* und *Cambridge English Dictionary* sowie das amerikanische *Merriam Webster* Aussprache-Audios (und vor allem die letzteren drei auch Transkriptionen) zunehmend in ihre Lemmata integrieren bzw. es bereits weitgehend getan haben, dass dies auch auf diverse Lernerwörterbücher oder zwei- und mehrsprachige Wörterbücher zutrifft und es zudem seit längerem auf die Aussprachevermittlung spezialisierte Internetplattformen wie die *Forvo (forvo.com)* gibt, so muss wohl damit gerechnet werden, dass der Benutzerkreis von orthoepischen Fachwörterbüchern mittel- und längerfristig deutlich schrumpfen wird.

Bereits jetzt fallen die deutschen und die englischen Aussprachewörterbücher im Vergleich zu modernen Universal- oder Lernerwörterbüchern etwa in Bezug auf die Zugriffsstruktur und die angebotenen Verlinkungen eher bescheiden aus. Ihre Zugriffsstruktur beschränkt sich im Wesentlichen auf die alphabetische Lemmaliste in den Printausgaben bzw. auf einfache Suchfunktionen in den digitalen Ausgaben. Erweiterte Optionen wie phonetische Suchmöglichkeiten sind nur in Einzelfällen und in begrenztem Umfang vorhanden. Auch die Verlinkungsangebote bleiben in den untersuchten Werken auf punktuelle Querverweise zwischen unterschiedlichen Aussprachevarianten oder Flexionsformen beschränkt. Eine systematische Vernetzung mit externen Ressourcen wie Thesaurus, Kollokationen oder (ggf. phonetisch relevanten) Grammatikhinweisen, wie sie in heutigen Universal- oder Lernerwörterbüchern üblich ist,¹¹ wird von den aktuellen Aussprachewörterbüchern nicht angeboten.

Sollten nun Universalwörterbücher wie die oben aufgezählten in Zukunft noch mit weiteren benutzerfreundlichen Funktionalitäten angereichert werden, etwa der Möglichkeit, die eigene Aussprache mit dem Standard-Audio des Wörterbuchs zu konfrontieren und online zu üben, dann lässt sich u. U. nicht vermeiden, dass eigenständige Aussprachewörterbücher schließlich auf ein Nischenphänomen reduziert werden, das hauptsächlich nur noch für Fachleute von Interesse bleibt.¹²

Endnoten

1. Zu einer kritischen Auseinandersetzung mit dem Wörterbuch siehe Lewis 2020: 147 ff.
2. In diesem Zusammenhang zählt DAWB auch einige konkrete Berufsgruppen auf, und zwar „Pädagogen, Mediensprecher, Schauspieler, Sprecherzieher, Rhetoriklehrer, Logopäden, klinische Linguisten“ (4. Umschlagseite).
3. Ausführlicher dazu siehe Nycz und Tęcza 2023: 168 ff., Nycz und Tęcza 2020b: 465 ff., Tęcza und Nycz 2016a: 384 ff.
4. Zur Kritik des *Oxford Dictionary of Pronunciation for Current English* siehe Picard 2004: 192.
5. Mehr zu formal-editorischer Beschaffenheit der deutschen Aussprachewörterbücher siehe Nycz und Tęcza 2020b: 76 ff., Tęcza und Nycz 2016b: 364 ff.

6. Vgl. <https://dad.sprechwiss.uni-halle.de/dokuwiki/doku.php/about/start>, Zugriff am 30.8.2025. Eingehender zu der DAB, deren Aufbau, geplanter Weiterentwicklung sowie besonders phono-didaktischem Potenzial siehe Ebel und Skoczek 2021: 39 ff., Skoczek und Ebel 2021: 73 ff.
7. Siehe <https://shop.duden.de/Duden-Das-Aussprachewoerterbuch/9783411040681>, Zugriff am 30.8.2025. Eine genauere Darstellung der Konstruktion und der Funktionsweise dieser Anwendung ist in Nycz und Tęcza 2020a: 85 ff. zu finden.
8. Vgl. <https://shop.duden.de/Duden-Das-Aussprachewoerterbuch/9783411914326>, Zugriff am 30.8.2025.
9. Zur Kritik des *Longman Pronunciation Dictionary* siehe Sobkowiak 2009: 191 ff., Lewis 2009: 238 ff.
10. Eine interessante und dabei noch in hohem Maße aktuelle Diskussion dieser Problematik wurde schon vor einem Jahrzehnt in Sangster (2015: 307 ff.) vorgenommen.
11. Vgl. z. B. das LDOCE, *Longman Dictionary of Contemporary English Online* (<https://www.ldoceonline.com>).
12. Wir möchten an dieser Stelle den beiden anonymen GutachterInnen für ihre wertvollen Kommentare und Anregungen danken, die die Qualität unseres Artikels erheblich verbessert haben.

Literaturverzeichnis

I. Deutsche Aussprachewörterbücher (mit Siglen)

- [DAWB] Krech, E.-M., E. Stock, U. Hirschfeld und L. Ch. Anders (mit Beiträgen von W. Haas, I. Hove und P. Wiesinger). 2009. *Deutsches Aussprachewörterbuch*. Berlin/New York: De Gruyter.
- [DU-23] Kleiner, St. und R. Knöbl (bearb. in Zusammenarbeit mit der Dudenredaktion). 2023. *DUDEN. Das Aussprachewörterbuch* (= DUDEN-Band 6; achte, komplett überarbeitete und erweiterte Auflage). Berlin: Dudenverlag.

II. Englische Aussprachewörterbücher (mit Siglen)

- [EPD] Jones, D. (Hrsg. von P. Roach, J. Setter and J. Esling). 2022. *Cambridge English Pronouncing Dictionary* (18. Aufl.). Cambridge: Cambridge University Press.
- [LPD] Wells, J.C. 2008. *Longman Pronunciation Dictionary* (3. Aufl.). Harlow: Pearson Education.
- [ODP] Upton, C., W.A. Kretschmar Jr and R. Konopka. 2001/2003. *Oxford Dictionary of Pronunciation for Current English*. Oxford: Oxford University Press.
- [RDP] Upton, C. and W.A. Kretschmar Jr. 2017. *The Routledge Dictionary of Pronunciation for Current English* (2. Aufl.). Abingdon: Routledge.

III. Sekundärliteratur

- De Schryver, G.-M. 2023. Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography* 36(4): 355-387.
DOI: <https://doi.org/10.1093/ijl/ecad021>
- Ebel, A. und R. Skoczek. 2021. Die Deutsche Aussprachedatenbank (DAD) — Potenzial für Deutsch als Fremd-, Zweit- und Muttersprache. *Korpora Deutsch als Fremdsprache* 1(2): 31-53.

- Lewis, J.W.** 2009. Wells, J.C. *Longman Pronunciation Dictionary* (3. Aufl.). *Journal of the International Phonetic Association* 39(2): 238-240.
DOI: <https://doi.org/10.1017/S0025100309003934>
- Lewis, J.W.** 2020. Clive Upton and William A. Kretzschmar Jr. (Editors), *The Routledge Dictionary of Pronunciation for Current English*, 2nd edn. 2017. Pp. xxxvii + 1561. ISBN 978-1-138-12566-7 (hardback); ISBN 9-78131545969 (ebook). *Journal of the International Phonetic Association* 50(1): 147-149.
DOI: <https://doi.org/10.1017/S002510031700038X>
- McKean, E. und W. Fitzgerald.** 2024. The ROI of AI in Lexicography. *Lexicography* 11(1): 7-27.
DOI: <https://doi.org/10.1558/lexi.27569>
- Nycz, K. und Z. Tęcza.** 2020a. On the Pronunciation Dictionaries of Contemporary German: Lexicographic Construction and Multimedia Components. *International Journal of Lexicography* 33(1): 73-89.
DOI: <https://doi.org/10.1093/ijl/ecz026>
- Nycz, K. und Z. Tęcza.** 2020b. On the Pronunciation Dictionaries of Contemporary German: The Concepts of Phonetic Standard and Differences in Specific Phonetic Issues. *International Journal of Lexicography* 33(4): 463-487.
DOI: <https://doi.org/10.1093/ijl/ecz032>
- Nycz, K. und Z. Tęcza.** 2023. Phonostilistische und regionale Variation der deutschen Aussprache in den orthoepischen Nachschlagewerken der dritten Generation. *Lexikos* 33: 166-183.
DOI: <https://doi.org/10.5788/33-1-1804>
- Picard, M.** 2004. Review of *The Oxford Dictionary of Pronunciation for Current English*. *Language* 80(1): 192-192.
DOI: <https://dx.doi.org/10.1353/lan.2004.0040>
- Ptasznik, B. und R. Lew.** 2025. Dictionaries versus AI Tools through the Eyes of English Majors. *International Journal of Lexicography* 38(2): 140-158.
DOI: <https://doi.org/10.1093/ijl/ecaf005>
- Sangster, C.** 2015. Pronouncing Dictionaries. Durkin, Ph. (Hrsg.). 2015. *The Oxford Handbook of Lexicography*: 292-309. Oxford: OUP.
- Skoczek, R. und A. Ebel.** 2021. German Pronunciation Database and Its Possible Applications in the Age of Homeschooling. *Lublin Studies in Modern Languages and Literature* 45(3): 117-127.
- Sobkowiak, W.** 2009. John C. Wells. *Longman Pronunciation Dictionary*. *International Journal of Lexicography* 22(2): 191-209.
DOI: <https://doi.org/10.1093/ijl/ecp003>
- Tęcza, Z. und K. Nycz.** 2016a. Deutsche Aussprachewörterbücher im Vergleich: Die Auffassung des phonetischen Standards und Divergenzen in phonetischen Einzelfragen. *Germanica Wratislaviensia* 141: 383-398.
DOI: <https://doi.org/10.19195/0435-5865.141.25>
- Tęcza, Z. und K. Nycz.** 2016b. Deutsche Aussprachewörterbücher im Vergleich: Die lexikographische Architektur und Prinzipien der Aufnahme lexikalischen Materials. *Germanica Wratislaviensia* 141: 361-382.
DOI: <https://doi.org/10.19195/0435-5865.141.24>

Where Tools Are Few: Constructing Limited-Word Bilingual Learner's Dictionaries in a Low-Resourced Community in Malawi

Ian David Dicks, *Ciyawo–English Dictionary Project (CEDP), Malawi;*
Tabor College, Adelaide, Australia
(iandicks@icloud.com)
(<https://orcid.org/0000-0003-0110-9596>)

Abstract: Many minority language communities in sub-Saharan Africa are required to navigate national education systems dominated by languages in which they are not largely proficient, and which they lack resources and tools to learn. When taken together with a low political will for resourcing minority languages, these become formidable barriers for learning and development (Matiki 2006: 244, 246; Prinsloo 2017: 20). Many minority languages are also not understood or spoken outside of these communities, and there are few resources to assist people to gain capacity in them, which results in large communication gaps. This paper examines how a team of non-professional Yawo lexicographers in Malawi worked to overcome these barriers in order to produce two limited-word bilingual learner's dictionaries (English–Ciyawo; Ciyawo–English) in book and smartphone application forms. More specifically, it explains how the team was trained, the development of a limited organic corpus from oral interviews as a method for supplementing an unbalanced corpus, and the importance of collaboration. The paper also articulates the linguistic and pedagogical theories that are foundational to the project, including that second-language learners develop second-language competence more efficiently through engaging with a curated headword list of high-frequency and high-relevance words (Nation 2022: 15-17; Bayetto 2018: 12).

Keywords: CIYAWO, POVERTY, BILINGUAL, LEARNER, DICTIONARY, EDUCATION, MALAWI, AFRICA, OXFORD, LEXICOGRAPHY, SMARTPHONE, APPLICATION, CORPUS, LANGUAGE

Opsomming: Waar hulpbronne skaars is: Die maak van beperkte tweetalige aanleerderwoordeboeke in 'n Malawiese gemeenskap met geringe hulpbronne.

Baie minderheidstaalgemeenskappe in sub-Sahara Afrika moet in nasionale onderwysstelsels werk wat oorheers word deur tale waarmee hulle onvertrou is en waarin hulle nie hulpbronne en gereedskap het om te leer nie. Wanneer dit saamval met 'n swak politieke wil om minderheidstale te befonds, word dit gedugte hindernisse vir leer en ontwikkeling (Matiki 2006: 244, 246; Prinsloo 2017: 20). Baie minderheidstale word ook nie buite hierdie gemeenskappe verstaan of gepraat nie, en daar is min hulpbronne om mense te help om taalvaardiger te word, wat tot groot kommunikasiegapings

lei. Hierdie artikel ondersoek hoe 'n span amateur-Yawo-leksikograwe in Malawi gewerk het om hierdie hindernisse te oorkom om twee tweetalige aanleerderwoordeboeke (Engels–Ciyawo; Ciyawo–Engels) met beperkte woorde in boek- en slimfoonformaat te produseer. Meer spesifiek word daar verduidelik hoe die span opgelei is, asook hoe 'n beperkte organiese korpus uit mondelinge onderhoude ontwikkel is as 'n metode om 'n ongebalanseerde korpus aan te vul, en die belangrikheid van samewerking. Die artikel neem ook die taalkundige en pedagogiese teorieë fundamenteel vir die projek onder die loep, insluitend dat tweedetaalleerders tweedetaalbevoegdheid meer doeltreffend ontwikkel deur gebruik te maak van 'n saamgestelde trefwoordlys van hoëfrekwensie- en hoogs relevante woorde (Nation 2022: 15-17; Bayetto 2018: 12).

Sleutelwoorde: CIYAWO, ARMOEDE, TWEETALIG, LEERDER, WOORDEBOEK, ONDERWYS, MALAWI, AFRIKA, OXFORD, LEKSIKOGRAFIE, SLIMFOON, TOEPASSING, KORPUS, TAAL

Introduction

The Ciyawo–English Dictionary Project was first conceived in the late 1990's by Australian Baptist intercultural workers working in Malawi who saw the need to communicate with the Yawo in Ciyawo, but struggled to locate adequate tools to help them do so. Although there are strong social, cultural and missiological reasons for wanting to learn Ciyawo, the decision to learn Ciyawo instead of Chichewa, the most dominant language in Malawi, was not an easy decision as there was a dearth of available language learning resources in Ciyawo at that time. The only Ciyawo grammar books and dictionaries in existence were between fifty and one hundred years old and out of print, with only a few collector's versions available in specialty book shops at exorbitant prices. The problem with these tools, however, was more than a lack of availability and cost. They also lacked elements and qualities that would enable second-language (L2) learners to grow their capacity to understand, speak, read and write Ciyawo efficiently.

Initially, to overcome the communication gap, Ciyawo word lists with English equivalents were created and utilised by intercultural workers. These proved to be inadequate as they did not provide definitions for multiple senses or example sentences to show how headwords are naturally used, as well as other useful information for L2 learners. Another problem with the initial Ciyawo word lists were that they were largely translated from English headword lists. These lists represented ideas that native English speakers wanted to talk about rather than the ideas the Yawo had about the world and their life situations.

In 2005, Ian Dicks, an Australian intercultural worker working with Baptist Mission Australia (formerly Global Interaction) and Steffi Nchembe, a Yawo language and culture assistant, began to conceive of developing a Ciyawo learner's dictionary. This dictionary would consist of Ciyawo headwords and other elements that L2 learners of Ciyawo need to grow their capacity to communi-

cate with the Yawo. The notion for a limited-word learner's dictionary was based on research regarding the needs of L2 learners, which shows that L2 learners grow their second language competence most efficiently by learning the high frequency and high relevance words first (Coxhead 2000: 213; Nation 2001: 17-18).

The size of language vocabularies varies greatly. For example, the Oxford English Dictionary (OED) claims to explain the meaning of 500 000 English headwords and phrases (<https://www.oed.com/information/about-the-oed>). There is, however, a vast difference between the number of words in a language and the number of words that the average native speaker knows and uses. Studies show that "native speakers of English typically learn between 15–20 000 word families"¹ (Webb and Nation 2017: 7, 14). It is also recognised that an L2 learner can communicate adequately with a much smaller curated vocabulary. Studies show that high-frequency words are of the greatest value to L2 learners (Webb and Nation 2017: 6). By learning the 2 000 highest-frequency word families in English, as designated by the British National Corpus (BNC), a person has the ability to understand 86–91 percent of discourse and text in a novel, 81–84 percent in a newspaper, 89 percent of a conversation, 90 percent in a film and 89 percent of a lecture (Webb and Nation 2017: 11-12).²

It has also been shown that it is less beneficial for L2 learners to learn additional high-frequency words from the third, fourth, fifth, etc. sets of 1 000 word families than it is to learn additional low-frequency words of high relevance (Nation 2001: 17-18). The Oxford 3 000 and Oxford 5 000 word lists have been developed on this theory. The headwords in the Oxford 3 000 list are selected using two criteria — frequency and relevance for learners (<https://www.oxfordlearnersdictionaries.com/about/wordlists/oxford3000-5000#>). Oxford University Press (OUP) identified the highest frequency words, drawing on the BNC and the OED corpus (Shaffer 2005: 136). To this high-frequency list they added words of 'high relevance' to English language learners by selecting high-frequency words from "a specially created corpus of Secondary and Adult courses published by Oxford University Press" (<https://www.oxfordlearnersdictionaries.com/about/wordlists/oxford3000-5000#>).

Dicks and Nchembe saw the value of learning high-frequency and high-relevance words for L2 learners of Ciyawo. Initially, this interest was in the concept of developing a high-frequency and high-relevance word list in Ciyawo for L2 learners of Ciyawo. However, through research it became obvious that there were not adequate resources available to do this. There was neither a Ciyawo general corpus or a high-relevance corpus from which to develop such a list, nor were there human resources to utilise the data if such lists emerged.

During exploration of the value of high-frequency and high-relevance words for L2 Ciyawo learners, Dicks and Nchembe became acutely aware of the Yawo community's language learning context. They noted the lack of resources and tools to help Yawo students learn English as a second language, which at the time was the language medium of instruction for upper-primary, secondary and tertiary education in Malawi.

The Yawo education experience

The Yawo of Malawi are disadvantaged within the Malawian education system, as they are constantly playing catch-up throughout their educational experience. This is due to Chichewa being used as the primary medium of communication in primary school. The Yawo are also disadvantaged when it comes to engaging with government social services in the areas of health, law and order, civil society and development initiatives, as well as religious activities. The majority of professionals who come to work in Yawo-dominant areas are not proficient in Ciyawo, nor are they able or grow their capacity in the language even if they wanted to, as there are few resources available to help them.

This situation is not unique to the Yawo. It is also true of many minority language communities in Africa and around the world. Trudgen (2000: 68) says that in many minority language communities there is "a crisis in being understood". What Trudgen (2000) says of the Yolŋu people, an Aboriginal community in Northern Australia, reflects in many ways the reality of the Yawo and other minority language communities throughout southern Africa. The Yawo sit in meetings, listening to Chichewa and/or English, but often do not fully understand these languages. Alternatively, they may understand but do not feel competent to engage in the dominant language being spoken and so they remain quiet. This results in a tangible isolation in their own country. Trudgen (2000: 69) similarly points out:

Poor communication stops people receiving almost all news or knowledge from outside their language and cultural domain. This includes day-to-day news and general information. It also includes what may well be life-saving information from health professionals. It stops them knowing what they are giving consent for, how to comply with medical instructions and how to intervene in their own health problems. In this way, poor communication directly impacts on high mortality rates.

In Malawi it is taken for granted that Chichewa is universally understood; however, this is not the reality. For example, it is common for the Yawo to return from a visit to the hospital not fully knowing their diagnosis or the treatment regimen. This is not because it was not explained, but because there is often a communication gap. The professional did not understand or speak even a modicum of Ciyawo and the patient was not conversant in English, or fully conversant or confident enough in Chichewa to ask questions.

Trudgen (2000: 77) states that the communication crisis operates two ways: firstly, professionals are unable to communicate with people in minority language communities about matters that impact their lives and well-being, even when discussing basic concepts using basic terms, and secondly, people in minority language communities struggle to share their needs and ideas with professionals in a dominant language.

The Yawo context in Malawi

There are 17 languages spoken in Malawi (Matiki 2001: 203). Four languages are considered official languages, namely Chichewa, Citumbuka, Chitonga and Ciyawo, together with English (<https://www.immigration.gov.mw/citizenship/list-of-languages-in-malawi/>). Chichewa and English were declared the official languages of Malawi in 1968, with 50,2 percent of the population claiming to be first-language (L1) speakers of Chichewa and 4,9 percent claiming to understand English at that time (Matiki 2001: 201). According to Matiki (2001: 215), English became the "instrumental, regulative, interpersonal and imaginative/innovative language" as it was the only language accepted in education, judiciary, parliament and the media. However, due to English not being widely understood or spoken, "Chichewa acts as the de facto national language and English is the de facto official language in the country" (Reilly 2019: 32).

Although it has been widely reported that sub-Saharan languages are not in danger of being supplanted by languages of the colonisers, they are at risk of being supplanted by other dominant languages in their countries. Brenzinger (in Kandybowicz and Torrence 2017: 2) says, "the most immediate threat to minority African languages are posed by other local languages or sub-national languages", which is the case with Chichewa in relation to the other languages in Malawi.

Ciyawo is a widely spoken language in South Eastern Africa. It is the first language of more than 3,2 million people in Malawi, Mozambique and Tanzania (Baldauf and Kaplan 2004: 85, 155). In Malawi, more than 2,3 million people identify as Yawo and they make up 13,3% of the population (Malawi Data Portal 2018).

Malawi, however, is one of the least developed countries in the world with half of the population living below the poverty line when measured against the Multidimensional Poverty Index, which assesses people's lives across four equal dimensions of Health and Population, Education, Environment, and Work.³ Significantly, under the Education dimension, low levels of literacy and issues with schooling contribute to the nationwide level of multidimensional poverty in the country. However, other indicators reveal the depth and breadth of poverty for the Yawo, as the incidence of multidimensional poverty is 65,7 percent for people in rural areas, which is where the Yawo predominantly live, compared to 20 percent in urban areas (National Statistical Office Malawi 2022: vi). Moreover, children and youth are among those most impacted by poverty in Malawi — 63,8 percent of children between 0 to 9 years and 61,8 percent of children aged 10 to 19 years record the highest incidence of poverty in the country (National Statistical Office Malawi 2022: vi). The significance of these statistics is amplified at a district level as Mangochi (78,4 percent) and Machinga (78,2 percent) are the districts in Malawi that have recorded the highest percentage of people experiencing multidimensional poverty. These are areas in which the Yawo are the majority ethnolinguistic community (National Statistical Office Malawi 2022: vi).

This data indicates that Malawi is suffering an education crisis. This is supported by other data that shows that the percentage of children unable to read at a level appropriate for their age is estimated to be 89 percent in Malawi (UNESCO 2024). Successive governments have attempted to address the education issue in various ways. For example, tuition fees were abolished in 1994 for all government primary school education. In 2018, the government also announced its intention to abolish fees for government-provided secondary education.

However, the education crisis in Malawi is more nuanced than merely poor literacy and numeracy skills, and truancy. The education system has largely benefited first language Chichewa speakers since 1968 when English and Chichewa were adopted as the national languages of Malawi (Trudell 2016: 44). In 1969, Chichewa was established as the medium of instruction for the first four years of primary school and English became the language of instruction from Standards 5–8 in primary school and throughout secondary school (Matiki 2001: 206). English is also a mandatory subject in which a pass mark is required for the Primary School Leaving Certificate examinations in Standard 8; the Junior Certificate examinations in Form 2, and the Malawi School Certificate of Education (MSCE) in Form 4 (Matiki 2001: 206).

In 1996, once multiparty democracy was established, the education policy was changed to recognise and structurally include other languages in the educational life of Malawi. It was hoped that this would lead to a new dawn in education for the speakers of minority languages. A new education policy was written that directed mother tongue languages to be used as the medium of instruction in the first four years of education (Reilly 2019: 33). English would continue to be used as the medium for instruction from Standard 5 onwards. Apart from a few pilot programmes, however, the Ministry of Education did not implement this policy in any significant way, upholding the status quo.

Although people who are ethnically Chewa only make up 40 percent of the Malawian population, Chichewa has become the *de facto* first language of several other large tribes in Malawi, including the Ngoni and the Lomwe. When all first language speakers of Chichewa are counted together, they represent approximately 70% of the population of Malawi, which is a large and influential block (Matiki 2006: 250).

In 2014, the language policy for education changed again. This time it stipulated that the language medium of instruction would be English from Standard 1 in primary school through to Form 4 in secondary school (Trudell 2016: 45). This policy was meant to bring more equity to the education system, however it has also not been widely supported or implemented.

There are several reasons for the lack of implementation of the minority language policy as well as other minority language initiatives, including the social significance of Chichewa in Malawi. Chichewa serves an interpersonal function and operates as a link between people with different cultural and linguistic backgrounds in Malawi (Matiki 2001: 212). According to Matiki (2001: 212), Malawians are more likely to use Chichewa than any other language to bridge linguistic

differences, including in the education system. Issues of low language competence among teachers in minority languages and English is also a factor, as is linguistic centrism within the Ministry of Education and District Education Departments, and the overall paucity of educational resources for the population (Matiki 2006: 246-247; Trudell 2016: 39). Malawi does not have sufficient resources for its current education system, let alone a "translanguaging" approach to education (Kamwendo 2000: 6; Reilly 2019: 34). The extent of the education crisis was made evident to the project team when visiting 60 primary and secondary schools in the Mangochi District during 2024 to conduct seminars on how to use a dictionary and dictionary smartphone application (app). One primary school on the Eastern shore of Lake Malawi had 1 363 students and 8 teachers, resulting in a student-to-teacher ratio of 170:1. Moreover, the high number of students means that in some schools, classes are conducted in the open, under trees, as there are insufficient classroom blocks. In regard to the paucity of language resources, at one large primary school in the Mangochi District the project team was shown a small closet containing approximately 20 books, which is vastly inadequate for a primary school of more than 2 000 students. The consequences of this are seen in the Malawi 2018 Census, which indicates that the percentage of Yawo completing primary, secondary and tertiary education is lower than any other ethnic group in Malawi (National Statistical Office Malawi 2019).

The impetus for a wider project

It was from this understanding of the Yawo education context that the Ciyawo–English Dictionary Project (CEDP) was created as a vehicle for developing two bilingual dictionaries for two different end-user groups. One dictionary, the Ciyawo–English Learner's Dictionary (CELD), would be developed primarily for L1 English speakers who want to learn Ciyawo for communicating with the Yawo in their language, thereby promoting dignity and equity in dialogue and conversation (Dicks 2022; 2024a). The primary end users of the CELD are non-Yawo, English-speaking adult learners. The other dictionary, the English–Ciyawo Learner's Dictionary (ECLD), would be developed for L1 Ciyawo speakers, primarily students who need to develop their competence in English for educational and employment purposes (Dicks 2018; 2024b). The target audience for the ECLD would be students in Standards 6–8 of primary school and Forms 1–2 of secondary school in Malawi.

The need for both of these dictionaries was seen as important and urgent. However, since the project team lacked adequate resources with which to develop the CELD — namely, a word list of high-frequency and high-relevance headwords in Ciyawo developed from a corpus, as well as the identification of key word senses and natural example sentences — the team decided to embark on developing the ECLD first, as resources were available to begin this part of the project.

Methods of development

It is well documented that certain resources are deemed essential for the construction of modern, bilingual learner's dictionaries. At centre stage is the corpus, consisting of machine-readable written texts for determining headword lists, sense hierarchies, definitions and example sentences that are to be included in the dictionary articles (Atkins and Rundell 2008: 53). However, even the best and largest corpus is of little use without people with educational, linguistic, lexicographic and technological expertise to design a dictionary for specific end users, and to construct it by utilising the data appropriately. Moreover, anyone who has been involved in the management of a dictionary project knows that they are multi-year, even decades-long projects that require significant funding for equipment, software, rent, salaries, consultancy, publishing, etc. A project is also dependent on oversight and government and community support to ensure that the end product is constructed to an acceptable standard and aligned with the ideals and needs of the community.

In the Yawo context of southern Malawi, virtually all of these resources were unavailable in any significant manner at the conception of the project, except for community support. Prior to 2005, there was not a recognised standard orthography for Ciyawo (Centre for Language Studies (CLS) 2005), and up until 2005, Ciyawo was written idiosyncratically in Malawi and other countries. Even after a cross-border initiative to standardise the Ciyawo orthography involving linguists from Malawi, Mozambique and Tanzania, some linguistic and education departments continued to write Ciyawo idiosyncratically, a decision that hampers the development of literature in Ciyawo to the present day (Banda et al. 2008).

A formidable barrier to creating a high-frequency Ciyawo word list was the dearth of machine-readable text available in Ciyawo with which to develop a Ciyawo corpus. The available texts in Ciyawo were more than a century old. One of these was Yohanne Abdallah's (1973) historical account of the Yawo, written in Ciyawo in 1909. There were also several collections of Yawo-sacred stories and proverbial tales documented and translated by Duff Macdonald dating from 1881 and 1882 (Macdonald 1881; 1882). The only recent texts of significance were a Yawo Bible translation produced by the Bible Society of Malawi, and portions of translated Scripture and Scripture teachings that were being produced by translation teams in Malawi and Mozambique in the late 1990's and early 2000's.⁴ The only other recent text was a collection of traditional Yawo proverbs and proverbial stories, *The Wisdom of the Yawo*, published by Ian Dicks in 2006 (Dicks 2006).

Also available, but not usable for corpus construction were Yawo word lists and older dictionaries, including Edward Steere's Collections for a *Handbook of the Yao Language* (1871), Alexander Heatherwick's *A Handbook of the Yao Language* (1902) and Meredith Sanderson's *A Dictionary of the Yao Language* (1954). Apart from not being in a machine-readable state, none of these dictionaries were constructed from a corpus and they contained archaic terms and definitions

that are no longer commonly used by the Yawo. These texts, however, would still be useful as refining tools once a corpus was developed, as would the *Mgopolela Malowe Jwa Ciyawo* (CLS 2013), a monolingual Ciyawo dictionary.⁵

Aware that the previous dictionaries were dated and largely developed without a corpus, the project team decided to do everything possible to produce a general corpus in Ciyawo from which to construct the CELD. However, there is much debate regarding what constitutes a valid corpus (De Schryver and Prinsloo 2000: 91). A corpus is commonly defined as,

a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research (Sinclair 2004).

Moreover, "a general corpus is typically designed to be balanced, by containing texts from different genres and domains including spoken and written, private and public" (De Schryver and Prinsloo 2000: 91).

Even the task of producing a general corpus was a challenge for an African language with so little available written text. Moreover, the written text that was available was largely limited to a single semantic domain — Christian Scriptures. Nevertheless, it has been suggested that "any corpus — however unbalanced — is to be a source of information and indeed inspiration. Knowing that your corpus is unbalanced is what counts" (Atkins et al. 1991). The unbalanced nature of the Yawo corpus is one fact that the project team were fully aware of.

The aim then became to create what De Schryver and Prinsloo (2000: 92) call a "structured corpus", which is the first step towards creating an organic corpus. Available machine-readable written texts in Ciyawo were gathered, mainly in the domain of Christian Scriptures and teachings. To these were added additional transcribed texts from recorded oral interviews in a variety of semantic domains. The recorded interviews were conducted in domains including but not limited to the following: (i) work (well digging, farming, food preparation, fishing, grafting, business, harvesting, hunting, piece work, making local fermented drinks, clay pot making, carpentry, preparing groundnuts, making hoes, making metal pots); (ii) travelling; (iii) family life (marriage, traditional rules for pregnancy, food, cooking); (iv) traditional medicine (role of a healer, use of charms); (v) Yawo traditional religion (practices and beliefs, sorcery, witchcraft); (vi) Islam (practices and beliefs, religious differences); (vii) historical events (war in Mozambique); (viii) health and well being; (ix) Yawo Games; (x) initiation (boys, girls, pregnancy, chieftaincy songs); (xi) stories (traditional Yawo tales, histories); and (xii) death (funerals).

The bank of text consisting of Scripture and Scripture teachings were mainly in the narrative genre and constituted 448 199 words of the corpus. In regard to the oral interviews, more than 190 interviews were conducted, adding more than 322 000 words to the corpus. The total number of words from all sources in the Ciyawo corpus came to 770 871.⁶

The corpus, however, was unbalanced, as 60 percent was made up of text from Scripture and Scripture teachings and 40 percent from transcribed oral interviews. Some people have raised questions regarding the unreliability of an unbalanced corpus, as well as the unreliability of a corpus developed from spoken data, especially unscripted conversations (cf. De Schryver and Prinsloo 2000: 94). However, it is "any corpus compiler's task to attempt to assemble a representative corpus for his/her specific needs" (De Schryver and Prinsloo 2000: 92). The specific need identified by the project team was to identify the 3 000 high-frequency and high-relevance words in Ciyawo, and to define them for an English-speaking audience in the CELD. Considering this, the oral nature of a large portion of the source material is seen as a strength, since it captures speech that is natural and frequent, and not at a higher and more rarified literary level. Ultimately, corpus design is constrained by what is available, and if there is no written text with which to create a more so-called balanced corpus, the compilers must rely on whatever means are available to them.

It has also been argued that different genres of speech events are of different value to a corpus, with unscripted conversations at one end and monologues and lectures at the other (De Schryver and Prinsloo 2000: 94). The project team used what is known as an "informal interview" approach when conducting oral interviews on selected domains, which placed the genre of speech at the monologic end of the spectrum. The people interviewed were those whom the community identified as having knowledge of a specific topic. The interviewers used various questioning methods to elicit extended responses, including the grand tour question, which allows a person to give an overview of the topic in response. This type of question was used predominantly because it encouraged a conversational tone and allowed participants to talk about what they considered important in regard to the topic, rather than focusing on what the interviewer deemed significant (Fetterman 2010: 40).

Once the interviewing and transcribing was complete, a corpus query tool, Wordsmith Tools 6.0, was used to process the text files and provide statistical information about the database. It also gave concordance lines showing how words are used in context, and enabled the tagging and assigning of word classes. Knowing the unbalanced nature of the corpus required the team to diligently sort, demote and remove lemmas that were overly represented, such as religious terminology. While Wordsmith works well as a query tool, it does not help in identifying low-frequency but high-relevance words. For this, the team consulted privately produced word lists and monolingual dictionaries, including the *Mgopolela Malowe Jwa Ciyawo* (CLS 2013), as well as older Ciyawo Dictionaries mentioned earlier, and the Yawo community.

The need for developing two dictionaries

The project was ambitious from the start. The aim was to produce two dictionaries for two different end-user groups. Some individuals inquired as to why

one bilingual dictionary could not be produced and then reversed or back-translated for the second language community. The reason this is not possible is because bilingual dictionaries are not like eggs — they do not flip easily, especially in the case of Ciyawo and English. This is because the high-frequency word lists are different for both groups of end users. However, even when actions or objects have gloss equivalents in both languages, the meanings applied to them can be different for each "languaculture" community (Agar 2002: 60). Even when the reality being defined seems to be ubiquitous to both language and culture communities — such as *bicycle/njinga*, *dog/mbwa*, *farmer/mlimi*, *to pray/kuswali* — it was found that L1 speakers of Ciyawo and English apply different meanings to these realities. This means that appropriate headword lists are required for each languaculture community, applying L1 definitions to each sense so that L2 learners can grow their understanding to include the L1 meaning of the headword rather than their own L2 cultural meaning.

The English–Ciyawo Dictionary

The purpose of the English–Ciyawo Learner's Dictionary (ECLD) is to assist L2 Ciyawo speakers of English in growing their English proficiency so they can navigate the Malawian education system, gain employment, and interact with L1 and L2 English speakers internationally.

Therefore, the headword list for the ECLD needed to consist of the 3 000 most high-frequency and high-relevance words for L2 learners of English. The project team did not have the ability to develop such a list. OUP was approached and asked if their 3 000 word list could be used, as well as definitions up to four senses, and example English sentences for these senses. While the agreement took time to arrange, OUP was helpful, generous, and accommodating throughout the entire process of interaction. Permission was granted to translate the OUP definitions into Ciyawo. For this task, a meaning-based method of translation was used. This method differs from a literal translation method in that it aims in "retelling, as exactly as possible, the meaning of the original message in a way that is natural in the language into which the translation is being made" (Barnwell 1986: 9).

The target audience for the ECLD is upper-primary and lower-secondary school students, unlike the *Oxford Advanced Learner's Dictionary* (OALD), which has been primarily designed for adult learners. This meant that a number of example sentences in the OALD 8th edition needed to be changed, as the ideas expressed in them were not easily recognisable by the Yawo end users (Hornby 2010).

To assist learners further, the translators sought to use only the 3 000 key English headwords in the example sentences. This enables Yawo learners to look up the words used in the English example sentences that they do not understand. Unfortunately, the OALD does not adhere to this principle strictly and includes words in a number of example sentences that are outside of the 3 000 key word list. To overcome this problem, permission was sought to use a consulting editor to locate replacement example sentences for all the senses

missing appropriate example sentences. This included senses in which the example sentences expressed an idea uncommon in a Yawo context, example sentences that used words outside of the Oxford 3 000 word list, and instances where an example sentence was not provided.⁷ Overall, the consulting editor provided 1 386 appropriate example sentences.

Training non-professional lexicographers

At the beginning, the project lacked people with the skills and abilities required to produce a dictionary, in particular lexicographers. Lacking available expertise, Dicks and Nchembe set to work learning the art and craft of dictionary construction. Dicks attended an intensive introductory course on lexicography at Stellenbosch University. Following this, Dicks and Nchembe researched practical lexicography using the works of Landau (2001), and Atkins and Rundell (2008). They also attended AFRILEX conferences to learn from people who were working on other dictionaries in the African context.

Drawing on this learning, research, and personal experience as learners of Ciyawo and English, Dicks and Nchembe created a style manual and included elements in the dictionary article that would be most helpful for L2 learners of both languages.

The dictionary articles in both dictionaries include the following elements:

- (1) Headword: The word that an L2 learner would be searching for.
- (2) Senses: This includes up to four key senses of meaning in the ECLD and up to three in the CELD.
- (3) Translation equivalent in L1 of learner: This applies if a lexical equivalent exists.
- (4) Definition: The definition of the headword in the learner's L1.
- (5) Example sentence in L2 of learner: This sentence or phrase is meant to show how the headword is used naturally according to the sense.
- (6) Translation of example sentence in L1 of learner: This would include the translation equivalent if a lexical equivalent exists.
- (7) Other tense forms: Verbs are shown in three tenses, namely present, continuous and past tense.
- (8) Other word forms: Other forms of the headword are shown if the word is an adjective.
- (9) Plurals: The plural form of a noun is given.
- (10) Part of speech: Ciyawo grammar terms are used in the ECLD, such as *lina* (noun), *msali* (verb), *mlondecesya* (adjective), *mjonjecesi* (adverb), *mpecesi* (preposition), etc. English grammar terms are used for parts of speech in the CELD.

After receiving feedback from end users on the ECLD that was published in 2018, several changes were made to the dictionary article for the construction of the CELD. A translation of the Ciyawo definition in English was added so that learn-

ers could read the definition in two languages. The number of senses was also reduced in the CELD from four to three key senses.

The lexicographical team

The original Yawo lexicographers were chosen primarily for their language capacity in both Ciyawo and English. They were also required to have completed the Malawi School Certificate of Education (MSCE), which is a recognised competence marker in English. Included in the first group of lexicographers and field testers employed were several primary and secondary school teachers who provided insight into the end users of the ECLD.

All of the lexicographers were provided with initial training, which included teaching on the nature of a dictionary, different types of dictionaries, the hallmarks of a learner's dictionary, the importance of understanding the end user: including their needs, abilities and their perspectives, the objectives of the Ciyawo–English Dictionary Project, and most importantly the principles and methods of defining words. This teaching covered various styles of defining words, including genus and differentia, synonym, typical, and complete sentence styles, as described in several practical guides to lexicography.

A key aspect of a learner's dictionary is the provision of an example sentence in which the headword is used appropriately according to the sense of the word. The ECLD used primarily example sentences from the OALD 8th edition and from the consulting editor, as already mentioned. However, for the construction of the CELD, the newly created Ciyawo corpus was the main source of these example sentences. It was not always possible to use the corpus, since available sentences and phrases were not always appropriate or clear. For the senses without appropriate example sentences, recorded role-plays were conducted by the team to capture the way a word is used in natural dialogue. This method was found to be helpful and also proved to be a lot of fun.

Collaboration

In a low-resource community collaboration is essential — not just for dictionary construction, but in every aspect of life. The Yawo realise this and have many proverbs that talk about the necessity of working together to achieve a purpose, such as 'Mtwe umpepe wangatwicila cipagala' ('One head does not carry a small thatched roof') (Dicks 2006: 96).

In many ways collaboration is what made it possible to undertake and complete this dictionary project. From the very beginning, collaboration was required between L1 English and L1 Ciyawo speakers, as neither community had the language abilities to construct a bilingual dictionary on their own. Collaboration extended further to virtually every aspect of the project. Within the lexicography office, people operated collaboratively: Knowing that individually people

were under-skilled, the lexicographical team always worked in pairs when creating and reviewing dictionary articles, and role-plays were conducted as a group.

Knowing that the team did not have the skills required for designing and constructing a single dictionary, let alone two, collaboration was sought with OUP, which resulted in permission to use designated parts of the OALD 8th edition. OUP is a large multinational company that could have declined these requests. Instead, to their credit, OUP helped this small project team to achieve its goals, while also protecting their intellectual property and brand.

The team collaborated with the CLS at the University of Malawi, Zomba. The CLS provided education and training on the Ciyawo Orthography. Furthermore, two Yawo-speaking academics from the CLS became consulting editors on the project. They contributed by reviewing and providing editorial insights on both dictionaries. The CLS also helped locate several Yawo graduates in linguistics who joined the CEDP lexicographical team.

Collaboration also occurred with the wider Yawo community. Yawo-speaking school teachers and school principals helped to facilitate field testing with students and teachers in primary and secondary schools. The Paramount Yawo chief, Ce Kawinga, visited the project on several occasions, as did other senior chiefs and local village headmen and women to understand the project and give encouragement. Senior Yawo chiefs also sent representatives for training as reviewers, and gave the team feedback and approval to publish the finished work.

The Malawi Ministry of Education provided support. This came after publishing the first English–Ciyawo Learner's Dictionary in 2018. The Minister of Education at the time, Hon. Bright Msaka SC, was the keynote speaker at the launch of the ECLD in Mangochi, and encouraged parents to purchase the dictionary for their children and the Ministry of Education to support the initiative. The recommendation of the Minister of Education facilitated collaboration with the Ministry of Education in Mangochi, enabling the third phase of the project to move ahead, which was to conduct seminars in primary and secondary schools in the Mangochi District on 'How to Use a Dictionary and Dictionary Smartphone Application'.

Project learnings

There were many lessons learnt from this project. The first is that dictionary construction takes longer than anticipated. The ECLD took 12 years from conception to publication, and an additional 4 years to complete and publish the CELD. Thereafter, it took an additional 2 years to release them as smartphone apps, add sound files, and conduct sensitisation workshops in primary and secondary schools in the Mangochi District on how to use them.

Another learning relates to the difference between the needs of dictionary end users and their ability to purchase one. Although Yawo students wanted and needed to learn English, very few had the financial means to do so even at a sub-

sidised price. The ECLD was published and printed in book versions in 2018 and the CELD in 2022. The initial print runs were small: Only 2 000 copies of the ECLD were printed, and of these 1 500 copies were distributed free of charge to primary and secondary schools in Yawo-speaking areas of Malawi, while only 500 copies were sold to the general public. For the CELD, 1 000 copies were printed, of which 512 copies were distributed without cost to schools and 488 copies were sold through various bookstores in Southern Malawi for between US\$5–\$10. Although the demand for the dictionary was great, the ability of individuals and the Malawian Government to purchase them was extremely low. Due to the poor economic circumstances of many Yawo and Malawians, books and dictionaries are rarely purchased because of their cost.

Dictionary smartphone apps

With the introduction of smartphones in Malawi and the high adoption of them by many households, even among the Yawo, it was decided to investigate publishing the ECLD and CELD in smartphone apps that could be operated without the need for cellular data once the dictionaries were downloaded. The 2018 census indicates that 51,7 percent of Malawian households own a mobile phone. Significant for this project, 47,1 percent of households in the Mangochi District own a cellular phone, which is the fifth highest level of cellular phone ownership of the 13 districts in Southern Malawi, apart from cities (National Statistical Office Malawi 2019). Moreover, the number of cellular connections continues to rise. In 2025 there are 13,2 million cellular connections, which represents 60,3 percent of the general population (Datareportal 2025).

Once again, the team lacked the technical expertise or financial resources to undertake such a venture independently. The idea of creating smartphone apps was initially discussed with OUP; however, their suggested pathway was technically and financially out of reach. To find a solution, the team investigated the Dictionary App Builder (DAB) programme, developed by the Summer Institute of Linguistics (SIL), an international Scripture translation organisation. The DAB enables non-professional app developers equipped with an appropriate lexicon data file to build a dictionary app at an extremely low cost. The two CEDP dictionaries were constructed using the dictionary writing system TshwaneLex, which provides multiple options for exporting a lexicon database. The DAB is a sophisticated programme that enables the construction of a dictionary smartphone app with all elements of a dictionary article that are beneficial to L2 learners.

The completed dictionary apps have several advantages over the paper form of the dictionaries. The app enables the end user to search for words in multiple ways, including without knowing how to spell the word fully. Morris (2021: 36) highlights this as a benefit, saying that "electronic dictionaries are structured in a completely different way and no longer rely on the alphabetical access structure as the only access, the outer alphabetical access structure loses its status as the default access structure". The DAB app also allows the end user to search

for words that are only partially known, and to search for and find words that are misrepresented orthographically, which according to Morris (2021: 39) is the sign of a good electronic dictionary interface.

The DAB app also displays each article on a single screen, unlike an e-reader version of a dictionary. This reduces the overpowering affect that a page of articles can have on a novice. The DAB app also allows the dictionary to have an audible dimension, which greatly adds to its pedagogical function. In the second version of the ECLD and CELD apps, sound files were added so that end users can hear the headwords spoken by L1 speakers of Ciyawo and English, which assists with their speech reproduction. Once again, the construction of the apps was made possible through collaboration with the SIL who provided the DAB programme free of charge. The only costs associated with publishing the dictionaries as smartphone apps, have been paying a technology consultant, who has directed and assisted the building of the apps for Android and iOS. Both dictionaries are now available on the Google Play Store and the Apple App Store free of charge.

Finally, the Ciyawo–English Dictionary Project has been a collaboration between the Dictionary making team and financial supporters, who were predominantly ordinary people in Baptist churches of Australia who wanted the Yawo to have more opportunities to complete primary, secondary and tertiary education. My own role, as senior editor and project manager has always been part-time and has been supported largely through people in Australian Baptist churches. Baptist Mission Australia have supported this project and overseen the finances and tax deductible giving, which has enabled this project to be completed.⁸

The cost of this project is hard to establish, as the costs associated with the senior editor and project manager role were not accounted for between 2008–2016. These roles were part-time and conducted in conjunction with other roles and responsibilities in Malawi. Overall, the senior editor and project manager roles were undertaken approximately two days a week for 14 years, and then one day a week for another 4 years.

It is estimated that USD \$385 000 was spent to develop, publish and print two dictionaries in book form, as well as to produce smartphone apps for both dictionaries so that they could become available on Android and iOS.⁹

Conclusion

Many minority language communities in sub-Saharan Africa are under pressure of becoming redundant as mediums of communication, as they are undervalued and under-resourced in language learning tools. The result is seen in low education outcomes, inadequate delivery of social services, a sense of disempowerment, and isolation. What has been shown in this article is that a team of non-professional lexicographers can produce language learning tools through collaborative

effort. There are many other minority language communities in similar situations to the Yawo. Realistically, dictionaries for these communities are unlikely to be profitable for large dictionary corporations in the short term. However, collaboration with minority communities will greatly benefit communities, making learning more equitable and assisting in closing the communication gap.

Endnotes

1. "A word family consists of a headword (for example, 'assume'), its inflects ('assumes', 'assumed', 'assuming'), and its derivations ('unassuming', 'unassumingly')" (Webb and Nation 2017: 7, 14).
2. Others have suggested that a better threshold for an L2 learner is 3 000 highest-frequency word families as this would increase their ability to understand 98 percent of most reading materials and 95 percent of vocabulary used in spoken discourse. Dang and Webb (2016) have suggested that high-frequency words lists should count lemmas and lemma headwords instead of word families as "lemma headwords are the most commonly used unit of counting both inside and outside the classroom" (Webb and Nation 2017: 11-12). Either way, there is agreement that learning high-frequency words is key to closing the communication gap for L2 learners.
3. There are 13 subcategories, referred to as indicators, under these four headings: Food Security, Drinking Water, Nutrition, Sanitation, School Attendance, Literacy and Schooling, Asset Ownership, Housing, Rubbish Disposal, Electricity, Child Labour, Job Diversity, Unemployment (National Statistical Office Malawi 2022: 7).
4. There were some other texts that were very old, and that were not accessible to be transcribed. The most significant texts in the Ciyawo language were mainly other Bible translations, including several New Testament books translated by Alexander Hetherwick in the 1890's, including *Utenga Wambone Wa Luka*, *Masengo Ga Wandumitume*, *Utenga Wambone Wa Marko*, *Utenga Wambone Wa Matayo*, *Utenga Wambone Wa Yohana*, *Achikalata Jua Paolo Jua Ndumitume Kwa Wa Korinti* and *Kalata Jua Paolo Jua Ndumitume Kwa Wa Rumi* (Houston 2022: 9). There were also some primers, produced by R.S. Hynde in 1892 and 1894 (Houston 2022: 6).
5. The CLS dictionary is a monolingual dictionary produced as a tool for the hoped-for initiative of vernacular education in Ciyawo, which was made policy by the Malawian government in 1996 but never developed beyond a few pilot projects.
6. There is conjecture as to what constitutes a word, especially in Bantu languages, as it depends on the orthographical style (De Schryver and Prinsloo 2000: 100). What takes three words in English, "I am going", is written conjunctively as a single word in Ciyawo — Ngwawula.
7. Lorna Morris (née Hiles) was a tremendous help to our project as she had experience as an editor working on the *Oxford South African Illustrated School Dictionary* (2008).
8. I want to particularly thank my colleagues at Baptist Mission Australia, formerly Global Interaction, who saw the value of this project and supported it over many years, particularly Mrs. Alison Nissley and Dr. John Davis, who oversee projects in the organisation and who championed it and encouraged me along the way, as well as the many people who supported this project financially.
9. This figure was calculated from SFI giving records between 2009–2025.

References

- Abdallah, Y.B.** 1973. *The Yaos: Chiikala cha WaYao*. Edited and translated by M. Sanderson. London: Frank Cass.
- Agar, M.** 2002. *Language Shock: Understanding The Culture of Conversation*. New York: Harper Perennial.
- Atkins, S., J. Clear and N. Ostler.** 1991. *Corpus Design Criteria*.
<http://www.natcorp.ox.ac.uk/archive/vault/tgaw02.pdf> [16 May 2025]
- Atkins, B.T.S. and M. Rundell.** 2008. *The Oxford Guide to Practical Lexicography*. Oxford/New York: Oxford University Press.
- Baldauf, R.B. and R.B. Kaplan.** 2004. *Language Planning and Policy in Africa: Botswana, Malawi, Mozambique and South Africa. Vol. 1*. Clevedon: Multilingual Matters.
- Banda, F., A. Mtenje, L. Miti, V. Chanda, G. Kamwendo, A. Ngunga, M. Liphola, C. Manuel, B. Siteo, S. Simango and M. Nkolola.** 2008. *A Unified Standard Orthography for South-Central African Languages: Malawi, Mozambique and Zambia*. Second edition. Monograph Series No. 229. Cape Town: CASAS.
- Barnwell, K.** 1986. *Bible Translation: An Introductory Course in Translation Principles*. Third edition. Dallas: Summer Institute of Linguistics.
- Bayetto, A.** 2018. *The New Oxford Wordlist Research Report*.
https://www.oup.com.au/__data/assets/pdf_file/0016/120841/SCHL_WORDLIST_Report-2018_FA_2-web.pdf
- Centre for Language Studies (CLS).** 2005. *The Orthography of Ciyawo*. Chileka: E+V Publications.
- Centre for Language Studies (CLS).** 2013. *Mgopolela Malowé jwa Ciyawo: Ciyawo Dictionary*. Blantyre: Dzuka.
- Coxhead, A.** 2000. A New Academic Word List. *TESOL Quarterly* 34(2): 213-238.
- Dang, T.N.Y. and S. Webb.** 2016. Making an Essential World List for Beginners. Nation, I.S.P. (Ed.). 2016. *Making and Using Word Lists for Language Learning and Testing*: 153-167. Amsterdam: John Benjamins.
- Datareportal.** 2025. *Digital 2025: Malawi*.
<https://datareportal.com/reports/digital-2025-malawi> [17 May 2025]
- De Schryver, G.-M. and D.J. Prinsloo.** 2000. The Compilation of Electronic Corpora: With Special Reference to the African Languages. *Southern African Linguistics and Applied Language Studies* 18(1-4): 89-106.
- Dicks, I.D.** 2006. *Wisdom of the Yawo People: Under the Elephant's Belly You Can't Pass Twice*. Zomba: Kachere.
- Dicks, I.D.** 2018. *English–Ciyawo Learner's Dictionary*. Mzuzu: Mzuni Press.
- Dicks, I.D.** 2022. *Ciyawo–English Learner's Dictionary*. Mzuzu: Mzuni Press.
- Dicks, I.D.** 2024a. *Ciyawo–English Learner's Dictionary*. Smartphone Application. Mzuni Press / Google Play Store / Apple App Store.
- Dicks, I.D.** 2024b. *English–Ciyawo Learner's Dictionary*, Smartphone Application. Mzuni Press / Google Play Store / Apple App Store.
- Fetterman, D.M.** 2010. *Ethnography: Step-by-Step*. Third edition. Los Angeles: SAGE.
- Hornby, A.S. (Ed.).** 2010. *Oxford Advanced Learner's Dictionary*. Eighth edition. Oxford: Oxford University Press.

- Houston, T.J.** 2022. Utenga Wambone — The "Good News": An Exploration of Historical Ciyawo Bible Translations and Linguistic Texts. *Studia Historiae Ecclesasticae* 48(3): 1-18.
- Kamwendo, G.H.** 2000. Interfacing Language Research with Policy: The Case of Language in Education in Malawi. *Nordic Journal of African Studies* 9(2): 1-10.
- Kandybowicz, J. and H. Torrence.** 2017. Africa's Endangered Languages: An Overview. Kandybowicz, J. and H. Torrence (Eds.). 2017. *Africa's Endangered Languages: Documentary and Theoretical Approaches*: 1-10. Oxford/New York: Oxford University Press.
- Landau, S.I.** 2001. *Dictionaries: The Art and Craft of Lexicography*. Second edition. New York/Cambridge: Cambridge University Press.
- Macdonald, D.** 1881. *East African Tales*. Edinburgh: William Blackwood and Sons.
- Macdonald, D.** 1882. *Africana; or, The Heart of Heathen Africa. Vol. 1*. London: Simpkin Marshall & Co.
- Malawi Data Portal.** 2018.
<https://malawi.opendataforafrica.org/gesjee/population-2018> [17 May 2025]
- Matiki, A.J.** 2001. The Social Significance of English in Malawi. *World Englishes* 20(2): 201-218.
- Matiki, A.J.** 2006. Literacy, Ethnolinguistic Diversity and Transitional Bilingual Education in Malawi. *International Journal of Bilingual Education and Bilingualism* 9(2): 239-254.
- Morris, L.H.** 2021. *A Model for a Comprehensive Electronic School Dictionary for South African Primary School Learners*. Unpublished Ph.D. Dissertation. Stellenbosch: Stellenbosch University.
- Nation, I.S.P.** 2001. *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.
- Nation, I.S.P.** 2022. *Learning Vocabulary in Another Language*. Third edition. Cambridge: Cambridge University Press.
- National Statistical Office Malawi.** 2019. *2018 Malawi Population and Housing Census. Main Report*.
<https://www.nsomalawi.mw/census/2018> [17 May 2025]
- National Statistical Office Malawi.** 2022. *The Second Malawi Multidimensional Poverty Index Report — Nov 2022*.
https://www.undp.org/sites/g/files/zskgke326/files/2023-08/Malawi-Multidimensional%20Poverty%20Index%20Report_0.pdf [17 May 2025]
- Oxford Learner's Dictionaries.** 2025. *The Oxford 3000 and the Oxford 5000*.
<https://www.oxfordlearnersdictionaries.com/about/wordlists/oxford3000-5000#> [17 May 2025]
- Prinsloo, D.J.** 2017. Africa's Response to the Corpus Revolution. Xu, Hai (Ed.). 2017. *Proceedings of the 11th International Conference of the Asian Association for Lexicography (ASIALEX 2017), 10–12 June 2017, Guangzhou, China: Lexicography in Asia: Challenges, Innovations and Prospects*: 20-31. Guangzhou, China: ASIALEX.
- Reilly, C.** 2019. Attitudes Towards English as a Medium of Instruction in Malawian Universities. *English Academy Review* 36(1): 32-45.
- Shaffer, D.E.** 2005. Review of *Oxford Advanced Learner's Dictionary of Current English* (7th ed.). *Korea TESOL Journal* 8(1): 135-141.
- Sinclair, J.** 2004. Corpus and Text: Basic Principles. Wynne M. (Ed.). 2004. *Developing Linguistic Corpora: A Guide to Good Practice*: 1-16. Oxford: Oxbow Books.
<https://users.ox.ac.uk/~martinw/dlc/> [16 May 2025]
- Trudell, B.** 2016. *The Impact of Language Policy and Practice on Children's Learning, Evidence from Eastern and Southern Africa*. Nairobi: UNICEF.
<https://www.unicef.org/esa/sites/unicef.org.esa/files/2018-09/UNICEF-2016-Language-and-Learning-FullReport.pdf> [16 May 2025]

<http://lexikos.journals.ac.za>; <https://doi.org/10.5788/35-2-2108> (Volle uitgawe / Full volume)
<https://lexikos.journals.ac.za>; <https://doi.org/10.5788/35-2-2089> (Project)

Where Tools Are Few: Constructing Limited-Word Bilingual Learner's Dictionaries 275

Trudgen, R. 2000. *Why Warriors Lie Down and Die*. Darwin: Aboriginal Resource and Development Services Inc.

UNESCO. 2024. *Malawi: Education Country Brief*.

<https://www.iicba.unesco.org/en/malawi> [17 May 2025]

Webb, S. and P. Nation. 2017. *How Vocabulary is Learned*. Oxford: Oxford University Press.

Generative Artificial Intelligence (GenAI) as Information Tool for Lexicographic Information Needs

Theo J.D. Bothma, *Department of Information Science,
University of Pretoria, South Africa*
(theo.bothma@up.ac.za)
(<https://orcid.org/0000-0001-7850-3263>)

and

Rufus H. Gouws, *Department of Afrikaans and Dutch,
Stellenbosch University, South Africa*
(rhg@sun.ac.za) (<https://orcid.org/0000-0002-3423-058X>)

Abstract: The focus of this paper is on GenAI, chatbots and some implications for lexicography and dictionary use. The use of GenAI as an information tool to provide information to end-users (readers) who have a specific information need when reading a text is discussed. GenAI could previously provide content similar to that of a dictionary but could not provide contextualised answers. This has changed and most chatbots now have two modes, a "search" mode and a "thinking/reasoning mode", i.e., it is able to argue logically about its different proposed meanings in context and tends to offer a solution. This feature is discussed at the hand of examples, and it seems as if the chatbots can now provide correct contextual meaning and logically motivate the choice of meaning in context, based on their critical analysis and thinking skills although it still "hallucinates" if it has no answer. Hereafter, the role of dictionaries and lexicographers in an AI-enhanced world is discussed. The traditional assignment of dictionaries to document the status and history of a language remains a very important function and needs to be encouraged, especially in environments with limited language resources. However, exploring new commercial ventures, incorporating latest technologies, would be essential to the future of the discipline and industry.

Keywords: CHATBOT, CONTEXTUALISATION, GENERATIVE ARTIFICIAL INTELLIGENCE, INTERFACE, HALLUCINATE, GENAI, INFORMATION NEEDS, INFORMATION TOOLS, MEANING, TEXT RECEPTION

Opsomming: Generatiewe Kunsmatige Intelligensie (GenKI) as inligtings-werktuig vir leksikografiese inligtingsbehoefte. Hierdie artikel fokus op GenKI, kletsbotte en sommige implikasies daarvan vir leksikografie en woordeboekgebruik. Die gebruik van GenKI as 'n inligtingswerktuig wat inligting moet verskaf aan eindgebruikers (lesers) met spesifieke inligtingsbehoefte wanneer hulle 'n teks lees, word bespreek. GenKI kon vroeër inhoud vergelykbaar met dié van 'n woordeboek verskaf, maar kon nie gekontekstualiseerde antwoorde verskaf nie. Dit het verander en die meeste kletsbotte het tans twee modi, naamlik 'n "soekmodus"

en 'n "dink/redeneermodus". Hulle kan hulle verskillende betekenisvoorstelle logies binne die gepaste konteks beredeneer en neig om 'n oplossing aan te bied. Hierdie kenmerk word aan die hand van voorbeelde bespreek, en dit lyk asof kletsbotte nou korrekte kontekstuele betekenis kan verskaf en die keuse logies binne die konteks kan motiveer, gebaseer op hulle kritiese analise en denkvaardighede, alhoewel hulle steeds "hallusineer" as hulle nie 'n antwoord het nie. Hierna word die rol van woordeboeke en leksikograwe in 'n KI-versterkte wêreld bespreek. Die tradisionele opdrag van woordeboeke om die status en geskiedenis van 'n taal te dokumenteer, is steeds 'n belangrike funksie wat aangemoedig moet word, veral in omgewings met beperkte taalhulpbronne. Die verkenning van nuwe kommersiële ondernemings, insluitend die nuutste tegnologieë, is egter noodsaaklik vir die toekoms van die dissipline en die industrie.

Sleutelwoorde: BETEKENIS, GENERATIEWE KUNSMATIGE INTELLIGENSIE, GENKI, HALLUSINEER, INLIGTINGSBEHOETES, INLIGTINGWERKTUIG, KLETSBOT, KONTEKSTUATILISERING, KOPPELVLAK, TEKSBEGRIP

1. Introduction

1.1 Background to artificial intelligence

Currently, artificial intelligence (AI) can be found in numerous and diverse environments. Many newspapers, popular tech journals, tech newsletters, academic articles, books, etc. discuss the uses of AI, with widely divergent opinions. On the one hand, there is a very positive approach, arguing that everybody should be AI literate and should actively use AI. On the other hand, there is a negative approach, arguing, *inter alia*, that AI heralds the end of humanity, the end of jobs and that it is bad for the environment. There are also various issues regarding ethics in the work environment, for students, school children and researchers. Problems of bias, plagiarism, privacy, lack of learning and "hallucinations" are contrasted with the apparent usefulness of this tool, even if it is used correctly. Current AI detection software is not able to differentiate unambiguously between AI generated content and authors' own work, as well as between mis-/disinformation, fake news and deepfakes. The potential positive impact of the AI revolution can, however, not be negated.

A significant aspect of GenAI has been the speed of its development. The first release of the first commercial chatbot was at the end of 2022. It was App- or browser-based and since then AI and GenAI have become common tools for users in all walks of life for the satisfaction of multiple types of information needs. It also enabled typical natural language search tasks from the GenAI interface. AI has also been embedded in many application software packages, social media software, system software (e.g., on smartphones such as iPhones and Samsung phones), and it has been embedded in operating systems, browsers, etc. An AI summary is often provided as the default answer to a Google search, presenting an AI Overview, followed by relevant links. Google's latest development, AI Mode, will have a further significant influence on the general AI-based

provision of information to all Google search users. AI Mode is characterized by Germain (2025) as follows:

People use Google Search five trillion times a year — it defines the shape of the internet. AI Mode is a radical departure. Unlike AI Overviews, AI Mode replaces traditional search results altogether. Instead, a chatbot effectively creates a miniature article to answer your question. As you read this, AI Mode is rolling out to users in the US, appearing as a button on the search engine and the company's app. It's optional for now, but Google's head of Search, Liz Reid, said it plainly when launching the tool: "This is the future of Google Search." [...] There is little doubt AI Mode is an impressive piece of technology. It deploys a "fan out method" where the AI breaks your question into subtopics and does multiple searches simultaneously. Google says this lets AI Mode recommend more diverse sources, produce deeper answers to more complex queries, dives deeper — and you have the ability to ask follow-up questions.

1.2 Terminology issues

When discussing AI, one has to be aware of some potential terminology problems. Industry tends not to specify the exact nature of the different AI developments and usages, but rather only employs the generic term *AI* to refer to all these developments. One should, however, strictly and clearly differentiate between:

- AI as an umbrella term for multiple technologies, such as NLP (natural language processing), neural networks, machine learning, deep learning, and symbolic reasoning.
- GenAI, based on large language models (LLMs)
- Agentive AI, which refers to autonomous agents
- General Purpose AI, Artificial General Intelligence (AGI), currently still an ill-defined industry term for the point at which human intellect is outpaced by algorithms
- Hybrid AI systems which combine different types of AI techniques to create more powerful and versatile AI solutions

The focus in this paper is on GenAI, chatbots and some implications for lexicography and dictionary use, with special reference to the end-users (readers) who have a specific information need when reading a text, i.e., a text reception information need.

Since the recent emergence of GenAI, its development has been phenomenal. This can be seen in the reliability of responses when using chatbots. Chatbots originally tended to "hallucinate", but the use of much larger training databases resulted in chatbots becoming more accurate. In this development, multiple iterations of chatbots from multiple software companies have been released. In addition, specialised chatbots for different environments have been developed, e.g., for general queries, coding, for research, text summarization and image, video and audio generation. Of major significance is the recent devel-

opment of two different modes, viz., to be the equivalent of a search engine, and the ability to "think" and to "reason," i.e., to argue about the validity of its answers.

1.3 AI and chatbots in lexicography

AI and chatbots have also been a focus in recent lexicographic research. The focus has mostly been on possibilities for dictionary compilation and the preparation of writing assistants. A prominent feature in modern-day lexicography is the user-perspective, postulating that "the dictionary user must be an important consideration in dictionary-making" (Hartmann 1989: 102). The main implication of the user-perspective is that the compilation of any dictionary should be preceded by the identification of the envisaged target user, the lexicographic needs of this user and their reference skills. However, this is often neglected. Wiegand (1977: 62) already urged that lexicography should strive towards a much closer relationship with the intended target user, and that this user should no longer remain the familiar stranger. Unfortunately, the user has yet again been neglected in recent developments regarding AI and chatbots in lexicography. Too little attention has been directed at the users and their needs and expectations.

Some recent research relevant to AI, lexicography and the linking between text, dictionaries and other information tools can be found in Bothma and Gouws (2022), De Schryver (2023), Bothma and Fourie (2024), Huete-García and Tarp (2024), Lew (2024), Fuertes-Olivera (2024), and Bothma and Fourie (2025).

1.4 Lack of context

Research directed at the use of AI and chatbots in lexicography has shown that GenAI could already provide content similar to a dictionary, but that it cannot provide contextualised answers. Consequently, the reader is still dependent on their own evaluation of the GenAI-provided content to determine the meaning of the word or phrase in context. This is to the detriment of the reader because even when sufficient context is provided in the prompt, the chatbot often provides only a single meaning/sense of a given word. The chatbot sometimes misunderstands the context provided in the prompt and could then easily provide an incorrect meaning. If then queried through a follow-up prompt why it chose a specific meaning, it could not provide any explanation. This causes doubt in the mind of the user and diminishes the user-friendliness and usefulness of the chatbot.

2. Empirical work

In this section, we provide a brief outline of our experimental methodology and the prompts that were used in the experiments.

2.1 Methodology

The experiments were done during March and April 2025. It was evidently not feasible to test and evaluate all chatbots, as the purpose of this paper is to show tendencies and development of chatbot capabilities (compared to 2024). The selection of GPT o3-mini was based on an AI Overview of the different versions of ChatGPT in Google that stated: "o3 and o3-mini: These are reasoning models designed for solving complex problems, especially in domains like research, strategy, coding, math, and science. They are good for structured, logic-driven thinking and layered reasoning" (March 2025). The choice for DeepSeek was because it was recently released, and according to popular reviews, was as powerful as or even more so than OpenAI's ChatGPT and other equivalents.

The choice of examples was based on previous experiences of the first author of this article with ChatGPT-3.5 and 4 during 2024, in which the chatbot "hallucinated" (i.e., provided incorrect answers), and when it was pointed out that the answers were incorrect, it could not provide any rationale or reasoning for those answers. In some of the current experiments, we used modified versions of two of the prompts of the 2024 experiments, supplemented with further examples to test the scope of the chatbots' ability to analyse prompts and reason.

In the examples and analyses below, GPT o3-mini provided the full analysis directly in its answer. DeepSeek, however, first provided its full thinking process (in a somewhat lighter font, but still easily accessible to the user), and then gave its detailed answer, based on its preceding thinking processes. Both chatbots also provided the number of seconds they "thought/reasoned" about a problem, in the current examples between 28 and 106 seconds.

2.2 Prompts

The full prompts we used to interact with the chatbots are given below and discussed in detail in Sections 2 and 3 of this article.

2.2.1 Prompt 1

"When asked about the meaning of the word 'Barbarossa' in the quoted text below, ChatGPT earlier suggested that it referred to the German army and their Barbarossa campaign in WWII. This was evidently wrong. The text reads: "'The Redbeard's German army approaches our northern borders and he offers me twenty thousand dinars!' He threw the scroll aside, where Imad al-Din picked it up and gazed at it earnestly, as if close scrutiny might reveal some previously concealed zeros. The German army. At the door of the war tent, Malek's heart dropped like a stone. Amid the glee of their recent successes, he had forgotten about Barbarossa and his advancing horde." If you disagree with ChatGPT's earlier explanation, why do you do so? What would be the correct meaning and why do you say so? Please provide references.

Prompt 1 tested whether the chatbot could analyse the example in its historical context, and identify a plausible historical context based on the available keywords. It also tested whether it could argue why one of the potential historical contexts is more likely than the other.

2.2.2 Prompt 2

"When asked about the meaning of the phrase 'Timothy's Book' in the quoted text below, ChatGPT earlier suggested that it referred to 'the first letter of Paul to Timothy, commonly known as 1 Timothy'. This was evidently wrong. The text reads: 'He always paused at the tomb of Prior Philip. It was the largest monument there. Philip, a twelfth-century monk, was a figure of legend, though not much was known about him. According to Timothy's Book, a history of the cathedral started in the Middle Ages and added to later, Philip had organized the rebuilding of the cathedral after it was destroyed in a fire.' If you disagree with ChatGPT's earlier explanation, why do you do so? What would be the correct meaning and why do you say so? Please provide references."

Prompt 2 tested whether the chatbot could analyse the example in a fictional historical context and argue from the formulation in the text of the novel, why the previous answer was illogical.

2.2.3 Prompt 3

"What are the different meanings of the word 'level'? Provide example sentences to illustrate the different meanings and include references."

With this prompt, we tested to what extent the chatbot could generate a valid dictionary article (or, at least, the outline of such an article) and illustrate each meaning with example sentences.

2.2.4 Prompt 4

"Can 'borrow' be used as a double transitive verb?" and "'Borrow' is often used as a double transitive verb in South African English. Is this an acceptable variant, or simply wrong?"

This prompt tested both senses and syntactic features of a word, and we tried to establish whether the chatbot could distinguish between regional variations in the use of words. For this example, we used Google AI overview (based on Gemini).

2.2.5 Prompt 5

"Provide a list of GenAI policies for research, teaching and learning in iSchools across the world (with references) and summarise the content of these policies"

and "Please provide a similar list and summaries for iSchools in the Global South, again with examples and references."

This is evidently not a text reception or purely lexicographic example. However, it does establish to what extent the chatbot is able to analyse a number of different, and typically unrelated, concepts to formulate a correct answer.

2.2.6 Prompt 6

"The sentence 'Time flies like an arrow' has multiple syntactic and semantic interpretations. List the different interpretations with, in each case, an analysis of the part of speech of each word in the sentence, together with references. Which is the most likely interpretation in your opinion?"

This is a complex problem with potentially multiple interpretations due to the ambiguity of the sentence. It requires the chatbot to use the potential meanings and morphological analyses of all the words in the sentence, based on linguistic and lexicographic criteria, to establish the most probable meaning of the ambiguous sentence.

Full details of the GPT o3-mini and Deepseek answers for this prompt are provided in the Addendum, available online (click [here](#)).

2.3 Examples of human-like communication

It is well known that chatbots communicate with users in narrative format, emulating human language characteristics. In this section, we provide examples of such characteristics under different headings.

2.3.1 Informal communication

Both chatbots in the experiments sometimes used informal language to create the impression of human communication, such as in the fairly extensive use of words such as "alright," "okay," "right," by both chatbots, as well as "Interesting!" and "Absolutely!" by DeepSeek.

2.3.2 Analysing prompts

The chatbots provided the users with an idea of how they went about analysing the problem and arriving at an answer, as in the following examples:

"Alright, let's break this down. [...] I need context from historical references [...]," "I initially thought about the WWII context, but no, that doesn't seem to fit," "This makes sense given the historical context of the narrative. I think I misunderstood earlier," "I need to provide sources," "I could refer to general works [...], but I must avoid fabricating references," "I remember several possi-

ble interpretations," "The sentence 'Time flies like an arrow' is famously ambiguous! Let's break it down." (GPT o3-mini)

"Okay, let's tackle this question [...] So I need to figure out why that answer was wrong and what the correct interpretation is," "Let me recall," "Let me read the quoted text again carefully," "The key here is [...]," "I need to explain [...]," "I need to structure all these aspects without overwhelming the user," "Let me make sure I cover every possible angle here." (DeepSeek)

2.3.3 Use of humour

DeepSeek sometimes included humorous statements in its answers, in the following examples:

"Buckle up — this will be a granular journey. 😊," "Let me know if you'd like to venture into string theory next ... or maybe just another sentence! 😊" (DeepSeek)

No such examples were observed in the current tests with GPT o3-mini.

2.3.4 Identifying contextual keywords

Chatbots identified the relevant keywords correctly, and used this in their analyses to substantiate their solutions, as in the following examples:

2.3.4.1 Prompt 1, the Barbarossa example

GPT o3-mini identified the relevant keywords correctly: "The text uses terms like 'dinar,' and names such as 'Imad al-Din' and 'Malek,' and it depicts scenes like a war tent — elements that point to a medieval or early modern Islamic context rather than the modern era."

DeepSeek also identified all the relevant keywords: "References to dinars (a historical currency), Arabic names (Imad al-Din, Malek), and a 'war tent' suggest a medieval Middle Eastern context, likely during the Crusades."

2.3.4.2 Prompt 2, the Book of Timothy example:

GPT o3-mini stated: "The passage explains that 'Timothy's Book' is a source stating that 'a history of the cathedral started in the Middle Ages ...' and goes on to recount that Prior Philip 'had organized the rebuilding of the cathedral after it was destroyed in a fire.' This language is typical of a chronicle or local historical account, not of a letter offering pastoral guidance."

DeepSeek stated: "The context is about Prior Philip, a 12th-century monk, and his role in rebuilding a cathedral after a fire. The text mentions "Timothy's Book" as a history of the cathedral started in the Middle Ages and added to

later. So, this book is a historical record of the cathedral's history, begun in medieval times and expanded over time."

2.3.5 Lexicographic and linguistic terminology

Chatbots could contextualise prompts in the specific subject/discipline, in the following examples from Prompt 6, "Time flies like an arrow." Both bots could provide the correct parts of speech for all five words in the sentence for all possible interpretations they listed. They also used relevant grammatical terminology, such as "plural," "imperative," "3rd person singular present" etc.

DeepSeek provided a fairly detailed answer in the first iteration (answer 1), but each time suggests to the user that it could give further analyses (answers 2–4), in which it analysed the following for each of the possible interpretations:

In answer 1, for each of the three interpretations, it provided Analysis, Meaning, Structure (Syntactic Tree) and a Reference;

In answer 2, the Syntactic Structure (Sentence Type and Phrase Structure), Semantic Analysis (Meaning and Plausibility);

In answer 3, Parts of Speech & Syntactic Roles (including Syntactic Tree (X-Bar Theory)), Semantic Analysis (Formal Semantics) and Pragmatic Factors;

In answer 4, it summarised the topics, viz., "phonetics, morphology, formal syntax/semantics, neurolinguistics, and even prosody." It then provided the following detail: Phonetic & Morphological Breakdown (Phonetic Transcription (IPA) and Morphology), Syntactic Analysis via Minimalist Program, Interpretation (Derivation and Features), Formal Semantics with Lambda Calculus (logical form), and, very briefly, Pragmatics & Gricean Maxims (Maxim of Relevance and Maxim of Manner), Psycholinguistic Processing (Eye-Tracking Predictions), Neurolinguistic Correlates, Computational Parsing Models, Historical & Cross-Linguistic Analysis, Prosodic Disambiguation, Quantum Linguistics (with a question mark and a smiley) and Final Conclusion at the Planck Scale. It concluded with the remark: "Let me know if you'd like to venture into string theory next ... or maybe just another sentence! 😊"

2.3.6 References

In each of the prompts, we requested the chatbot to provide references. We checked that the references actually existed and were relevant; in contrast with earlier bots, they did not "hallucinate," and the references existed. In some cases, the references were fairly broad and did not refer to the actual example, but provided more detail about the "setting" of their answer in a broader context, as is evident from the analyses of prompt 6, where the chatbots provided references to a number of well-known linguistics textbooks, or historical reference sources, in the case of prompts 1 and 2 (DeepSeek); GPT o3-mini tended to use more popular sources, e.g., to *Wikipedia* (in the case of prompt 1), and *Encyclopaedia Britannica* (as well as a specialist history book) (for prompt 2).

For the less successful (prompts 3 and 4) and unsuccessful (prompt 5) examples, there were some irrelevant and "hallucinated" sources, as detailed in Sections 3.2 and 3.3.

3. Discussion of examples

In this section, we discuss the successful, less successful and unsuccessful examples of chatbot consultations.

3.1 Successful

Both chatbots dealt successfully with prompts 1, 2 and 6.

Both chatbots were remarkably successful in analysing prompts 1 and 2. Both were able to identify relevant keywords in both examples and use these keywords to guide their interpretations. They argued convincingly that the keywords in prompt 1 pointed to a medieval setting, and not to a modern setting, which excluded the Barbarossa campaign of World War II. In prompt 2, both pointed out that the context of the Book of Timothy excluded a biblical interpretation, viz. a reference to the Pauline letter of 1 Timothy. For both prompts, both chatbots were therefore able to place the texts in a correct period and point out that the later interpretation (World War II) and earlier interpretation (a biblical text) were contextually not correct.

3.1.1 Prompt 1

GPT o3-mini provided the following summary answer for prompt 1:

"The earlier explanation was incorrect because it ignores several key contextual clues in the quoted text. Here's why:

"1. Historical Setting:

"The text uses terms like "dinar," and names such as "Imad al-Din" and "Malek," and it depicts scenes like a war tent — elements that point to a medieval or early modern Islamic context rather than the modern era. In contrast, Operation Barbarossa (the German invasion of the Soviet Union in 1941) is a 20th century event, making that explanation anachronistic.

"2. Meaning of "Barbarossa":

"The term "Barbarossa" comes from the Italian *barba rossa*, meaning "red beard." This nickname was historically associated with figures such as Frederick Barbarossa, the 12th century Holy Roman Emperor known for his red beard. In literary or historical texts set in medieval times, using "Barbarossa" would more likely evoke the memory or legend of such a figure rather than referring to a modern military campaign.

"3. Narrative Distinction:

Notice that the text separately mentions "The Redbeard's German army" and later "Barbarossa and his advancing horde." This implies that the invading force led by Barbarossa is a distinct entity from the other German army mentioned. In a WWII context, "Barbarossa" is synonymous with the Nazi operation, not as a separate leader or faction. The separation here reinforces that the name is being used in a medieval sense — as a proper noun evoking a legendary invader.

"In summary:

"Given the language and imagery, the correct interpretation is that "Barbarossa" in this text is a reference to a medieval invader — likely evoking the historical persona of Frederick Barbarossa (or a similarly legendary figure) — and not the WWII campaign. This interpretation fits the cultural and historical setting of the narrative much better than the modern association.

"For more details on the historical figures and events associated with the name, see:

- [Frederick Barbarossa – Wikipedia](#)
- [Operation Barbarossa – Wikipedia](#)"

3.1.2 Prompt 2

DeepSeek provided the following summary answer for prompt 2:

"The phrase "Timothy's Book" in the quoted text does **not** refer to the biblical First Epistle of Paul to Timothy (1 Timothy). Here's why:

"1. Contextual Discrepancy:

- The text explicitly describes "Timothy's Book" as "a history of the cathedral started in the Middle Ages and added to later."
- **1 Timothy**, by contrast, is a New Testament letter written in the **1st century AD** addressing pastoral and ecclesiastical issues. It has no connection to medieval cathedral histories or chronicles.

"2. Fictional or In-Universe Reference:

- The passage appears to be from **Ken Follett's *The Pillars of the Earth***, a historical novel centered on the construction of a medieval cathedral. In this context, "Timothy's Book" is a fictional chronicle created by Follett to serve as an in-universe historical record.
- The novel often references fictional documents to ground its storytelling in a sense of historical authenticity. For example, Prior Philip (a key character) is depicted as a legendary figure whose deeds are recorded in this fictional "Timothy's Book," which spans generations of cathedral history.

"3. Why ChatGPT's Initial Answer Was Incorrect:

- ChatGPT likely conflated the name "Timothy" with the biblical epistle, missing the **medieval and fictional context** of the text. The mention

of a cathedral's history being compiled over centuries (post-12th century) rules out a 1st-century biblical text.

"4. References:

- In *The Pillars of the Earth*, Follett uses fictional chronicles and manuscripts to enrich the narrative. "Timothy's Book" is one such device, akin to the *Kingsbridge Chronicles* (the series to which the novel belongs).
- The novel itself is set in the 12th century, and the fire/reconstruction of the cathedral aligns with historical events like the real-life destruction of medieval churches (e.g., the 1174 Canterbury Cathedral fire).

"Conclusion:

"Timothy's Book' is **not a biblical text** but a fictional medieval chronicle within the novel's universe, documenting the cathedral's history. This distinction highlights the importance of **context** when interpreting references in historical or literary works."

It is evident that the reasoning and argumentation are similar to the thinking and reasoning processes of a human who is interested in the historical details of the text.

3.1.3 Prompt 6

The discussions of Prompt 6, about the different meanings of the sentence "Time flies like an arrow," were equally very successful. Both bots analysed the sentences in detail, providing part of speech, syntactic and semantic analyses. Both also provided three different possible interpretations, adapting the part of speech, syntactic and semantic analyses to reflect the specific potential interpretation. They furthermore provided a motivation for their specific choice and motivated this in detail. Both chatbots provided a number of relevant references, all of which existed (i.e., no "hallucinations").

The full answers of both chatbots are provided in the Addendum, available online (click [here](#)), and make for quite interesting reading.

GPT o3-mini introduced and discussed its reasoning process as follows:

"The sentence 'Time flies like an arrow' is famously ambiguous! Let's break it down.

"Interpretation 1: It can mean that time quickly passes, much like an arrow's flight.

"Interpretation 2: 'Time flies' might refer to insects (as a noun), with 'like an arrow' being a comparison.

"There are multiple ways to interpret the parts of speech here: 'flies' as a noun (the insects) or verb (the action), and 'like' as a comparison. It's a classic case of ambiguity.

"I remember several possible interpretations."

This is followed with a detailed discussion of more than 800 words. For a full transcript, see the Addendum, available online (click [here](#)).

At the end of each answer for this prompt, DeepSeek asked whether further, more detailed analyses were required, and provided further details that may potentially be relevant for a specialist, or, citing some of the potential motivations and contents.

DeepSeek introduced and discussed its reasoning process as follows:

Answer 1:

"Okay, so I need to figure out the different interpretations of the sentence 'Time flies like an arrow.' I remember that this sentence is often used as an example of syntactic ambiguity. Let me start by breaking down each word and considering their possible parts of speech;" "First, the straightforward interpretation is probably 'Time passes quickly, similar to how an arrow flies.;" "But the user mentioned there are multiple interpretations. Let me think about other possible ways to parse this;" "Wait, another possibility ...;" "Wait, this is getting complicated;" "But that's a stretch ...;" "I need to make sure I'm not missing any other interpretations. Let me think again."

Answer 2:

"The user might be interested in the linguistic theories behind ambiguity. Chomsky's deep vs. surface structure comes into play here. The same surface sentence has different deep structures based on parsing. Radford's analysis of syntactic ambiguity would support this. Also, mentioning lexical ambiguity with homonyms is important;" "Also, consider the user's possible deeper needs. They might be studying linguistics, preparing for an exam, or just curious about language. By elaborating, I should provide enough detail to satisfy academic curiosity but keep it understandable. Avoid jargon where possible, or define terms if used," "Let's delve deeper into the syntactic and semantic ambiguities of 'Time flies like an arrow, expanding on linguistic theories, structural analyses, and contextual plausibility."

Answer 3:

"Okay, the user has already asked for a more granular analysis after I provided a detailed breakdown. They seem really interested in the linguistic aspects of the sentence 'Time flies like an arrow. Let me make sure I cover every possible angle here," and "Also, are there other possible interpretations I haven't considered? Maybe 'like' as a conjunction, though that's less likely. Or 'arrow' as a modifier? Probably not. But I should check;" "I need to make sure each analysis is thorough, referencing linguistic theories, providing syntactic trees with proper notation, and discussing semantic plausibility. Also, check if there are any other references or studies that

discuss this specific sentence — maybe classic examples in ambiguity literature."

Answer 4:

"Let me think about how to go deeper. First, maybe I should look into phonetics and morphology. But the sentence is in English, and the words are straightforward." It then briefly mentions "homonyms and their phonetic representations," "syntactic theories beyond X-Bar. Maybe using Minimalist Program structures, discussing movement or feature checking;" "Semantic roles could be expanded. Theta roles were mentioned;" "Pragmatic aspects could include Gricean maxims in more depth;" "computational linguistics;" "psycholinguistics;" "historical linguistics;" "morphological analysis;" "lexical semantics;" "neurolinguistics;" "prosody." It concludes its reasoning as follows: "I need to structure all these aspects without overwhelming the user. Organize them into sections with clear headings, use linguistic notation where applicable, and reference theories and studies for each point. Ensure that each detail adds value and connects back to the main interpretations. Check for coherence and avoid redundancy from previous answers. Highlight how each tiny detail contributes to the overall understanding of the ambiguity."

The four answers of DeepSeek result in a detailed discussion of approximately 6,000 words. For a full transcript, see the Addendum, available online (click [here](#)).

3.2 Partially successful

The answers to prompts 3 and 4 were less successful, primarily due to poorly structured groupings and "hallucinations" in the answer to prompt 3, and inconsistencies in the answers to prompt 4.

3.2.1 Prompt 3

In prompt 3, the chatbots were required to construct the equivalent of a dictionary article for the meanings/senses of a single word, *level*, as well as provide references.

GPT o3-mini introduced its answer as follows: "The word 'level' is a versatile term in English, used in different contexts with varying meanings. Below are the main meanings of 'level', each with examples and references." In each case, an example paraphrase or sentence from a dictionary is provided, in this case, the *Oxford English Dictionary (OED)*, *Merriam-Webster*, *Cambridge Dictionary*, *Collins English Dictionary* and *Longman Dictionary*. The examples are in all cases provided in quotation marks, which creates the impression that they are taken verbatim from the specified dictionary, but in none of the examples this

is the case — chatbot in each case provided a paraphrase. In addition, the proper dependencies between meanings/senses (as is evident in all the cited dictionaries) were not maintained. In this case, the chatbot is therefore a less dependent source of meanings/senses.

The chatbot also offered the option that it could provide examples from only one specific dictionary: "Let me know if you'd like definitions from a specific dictionary like OED or Merriam-Webster with direct quotes." We asked it to provide an answer from only the OED. None of the examples in this answer, however, were direct quotes from the OED — unless it referred to one of the other dictionaries of English published by Oxford, and not the "*The historical English dictionary*," typically known as "*Oxford English Dictionary*," and abbreviated as "OED" (<https://www.oed.com/>).

This is a less successful answer than could be obtained by typing "level meaning" in a Google search bar, which provides a comprehensive article with examples, synonyms etc.

3.2.2 Prompt 4

Prompt 4 (a query in the Google search bar, the answer of which is an AI Overview based on Gemini) is included to highlight confusion which sometimes occurs in GenAI answers. We asked whether the verb "borrow" could be used as a double transitive verb. In a first answer, it said: "Yes, the verb 'borrow' can be used as a double transitive verb, also known as a ditransitive verb." The rest of the explanation and examples seemed counterintuitive. We therefore repeated the question about an hour later (to check consistency in its responses), and the answer it gave was: "No, the verb 'borrow' is not typically used as a double transitive verb. While 'borrow' does take a direct object (what is being borrowed), it doesn't usually take an indirect object (to whom or for whom)." Since "borrow" is often used as a double transitive verb in (informal) South African English (i.e., as an equivalent of "lend"), we asked whether this usage is an acceptable variant in South African English, or simply a mistake. The answer read as follows: "While the use of 'borrow' as a double transitive verb (e.g., 'Can I borrow you the car?') is common in South African English, it's generally considered a nonstandard usage in formal English. It's often a loan translation from Afrikaans, where 'leen' (lend) and 'leen by' (borrow from) are used in a way that translates directly to 'borrow' in the double transitive construction. While it's widely understood and accepted in informal contexts, it's best avoided in formal writing or speech." This is the expected answer (which may confirm the authors' biases). Interestingly, this possible meaning is not listed in the *Dictionary of South African English* (DSAE, <https://www.dsae.co.za/>). The correct/incorrect reply from Google Overview AI highlights the importance of evaluating any answer GenAI provides, and of triangulating it, especially if it is counterintuitive.

3.3 Unsuccessful

Prompt 5 is not a lexicographic example, and is included to show that, even though chatbots have improved considerably since earlier versions, they still hallucinate if their training data does not contain the required data to answer the question.

3.3.1 Prompt 5

We used two prompts for this question. In the first prompt, we required the chatbot, in this case, DeepSeek, to be able to identify the different iSchools in the world (see the iSchools website at <https://www.ischools.org/>) and then find on the websites of the iSchools any policies for the use of GenAI in research, teaching and learning. The answer DeepSeek provided seemed to be correct, but very general. It provided references from 2023, however, without URLs to enable the verification of the existence of the sources, or more information. Some of these documents could be found, but in most cases, documents with the titles that were listed, were not available. It also provided a number of references under the heading "Hypothetical References," which is very strange.

We also needed information on the same topic, restricted to iSchools in countries in the Global South. This would imply that DeepSeek should be able to understand which countries the Global South consisted of, to match this list with the list of iSchools worldwide and then find the relevant policies. It was unfortunately not successful at all. We knew that, at that stage, there were four iSchools in Global South countries, one each in Brazil and Indonesia, and two in Africa. DeepSeek identified six iSchools in Global South countries, five of which were wrong and only one which was correct. It again provided incorrect or non-existent references and again added "Hypothetical References." The discussions were very general and none of the content could be verified.

These two examples again highlight the importance of evaluating and triangulating any answer GenAI provides.

4. GenAI versus dictionaries

The preceding discussion and examples show that GenAI cannot fully replace "conventional" dictionaries — at this stage. Access to actual dictionary sources provides better assistance, e.g., a search in online dictionaries that are available via open access, or even a Google search that provides "Definitions from Oxford Languages," results in retrieving better and more comprehensive definitions, as well as additional lexicographic information on demand. Currently, GenAI cannot be regarded as an authoritative lexicographic source, even when it claims to cite online dictionaries (as in prompts 3 and 4); it can, however, be an additional information tool for more complex information needs, as in the cases of

prompts 1, 2 and 6. In addition, it still "hallucinates" if it does not have an answer, as in the case of prompt 5. The potential unreliability of any GenAI-generated answer again emphasises the importance of evaluating any such answers and triangulating the results with other reputable (or curated) sources.

GenAI-enhanced tools are capable of responding to lexicographic information needs and to combine the obtained data with additional resources, constituting a network of information resources, as is evident in prompts 1, 2 and 6.

Irrespective of the advantages and disadvantages of AI, one has to accept that AI and GenAI are here to stay, at least for the foreseeable future. One cannot ignore these developments and should rather strive to understand what it can and cannot do in order to embrace it intelligently and ethically.

5. Suggestions for lexicographic research

5.1 Traditional lexicographic research

A focus in "traditional" dictionary research has always been, and continues to be, important, as dictionaries can best document and display the status and history of a language. This is witnessed in dictionaries such as the OED (*Oxford English Dictionary*, <https://www.oed.com>), WAT (*Woordeboek van die Afrikaanse Taal*, <https://www.wat.co.za>), the DSAE (*Dictionary of South African English*, <https://www.dsae.co.za>) and many others. These endeavours need to be encouraged, especially in environments with limited language resources. The lexicography of the African languages in South Africa could also benefit from such an approach. The dynamic nature of language demands that dictionary definitions should continuously be refined.

No single dictionary can be everything for everyone, but in the absence of an established typological infrastructure, existing dictionaries should increase their scope beyond the traditional typological boundaries so that they can present a wider linguistic coverage to their target users. One option the lexicographer has, is to plan and compile multifunctional online dictionaries with added functionalities, e.g., language-learning exercises, lexicotainment, etc. They could also offer access to additional lexicographic and non-lexicographic data, and/or links to such data, even to data in dictionary-external sources. This demands the appropriate interfaces and interface design that allow the filtering of data as well as drilling down and retrieving information on demand. Interface design is an exceptionally important aspect for the development of a successful online dictionary, and typically requires a number of specialists to collaborate, including lexicographers, UX (user experience) designers, programmers/computer scientists, information scientists and a number of others, as well as users to test the efficiency of the proposed interface.

When planning the interface design, it would be helpful to be reminded of Ben Shneiderman's Visual Information-Seeking Mantra (1996): "Overview first, zoom and filter, then details on demand." This should be understood as:

- "Overview: Gain an overview of the entire collection"
- "Zoom: Zoom in on items of interest"
- "Filter: Filter out uninteresting items"
- "Details-on-demand: Select an item or group and get details when needed"

To this could be added "Relate", "History" and "Extract", realised as:

"Relate: View relationships among items."

"History: Keep a history of actions to support undo, replay, and progressive refinement."

"Extract: Allow extraction of sub-collections and of the query parameters."

These aspects of interface design will not be discussed in this paper. Existing online dictionaries already offer an interface design that can guide the user within a comprehensive article to both a number of restricted articles, cf. Gouws (2022: 120), in which various aspects of the treatment of the lemma are presented or to dictionary-external data like other dictionaries in the same portal, or even to Google and other search engines. An option that modern-day lexicographers should consider is to integrate access to a chatbot, enabling dictionary users to retrieve complementary information to enhance the quality and extent of their dictionary consultation experience. This integration of GenAI and lexicography requires, among others, the appropriate interface design and data distribution structure.

5.2 Lexicographic research and technology

Lexicographic databases, their compilation and use are focal areas in modern-day lexicographic research. These databases and the lexicographic products relying on them can be improved with more attention paid to the concepts of granularity and (re)usability. Granularity is required for the fine-grain filtering and display of lexicographic data, whereas (re)usability should play a pertinent role in deriving different lexicographic products, including dictionaries for specific user groups from the data on offer. Different applications, e.g. translation tools, writing assistants and tools that help users to improve their writing can employ the same data but applied in different ways.

Lexicographic data becomes the gold standard in the planning and compilation of lexicographic products, and the data assessment, distribution and presentation need to be done in such a way that different target groups can use different tools to retrieve the information they require in specific extra-lexicographic situations. The optimal use of these data can be improved if lexicographers, AI experts and other collaborators can successfully explore and license new commercial ventures, incorporating the latest technologies, including AI and GenAI. Such ventures could also elevate the level of contextualized linking for text reception.

6. Conclusion

Based on the limited number of examples that have been examined, and contrary to earlier research, it seems as if the latest versions of GenAI chatbots can often provide correct contextual meaning. It can logically motivate the choice of meaning in context, based on its critical analysis and thinking skills, typically associated with humans.

However, the reliability of the results of GenAI answers varies, covering the full spectrum from very reliable to complete hallucinations. A mixture of reliability criteria prevails. As a result, the reader remains responsible to evaluate the answers. In case of uncertainty, they need to check other resources in order to verify and triangulate the results.

When expanding the information retrieval structure of a dictionary to allow users access from any point in a given dictionary to one or more chatbots, lexicographers need to make their users aware of the fact that GenAI provides non-curated data that might not always live up to the correctness, appropriateness and validity of the information they can retrieve from the curated data in dictionaries.

However, good lexicography demands interdisciplinary collaboration. An interactive relation between lexicographers and developers of GenAI could be beneficial for both groups. Lexicography could benefit by improved versions of GenAI becoming sophisticated lexicographic information tools and GenAI could benefit from the expertise in lexicography regarding the user-perspective, information needs, lexicographic functions, appropriate contents and relevant structures to distribute and access data.

7. Addendum

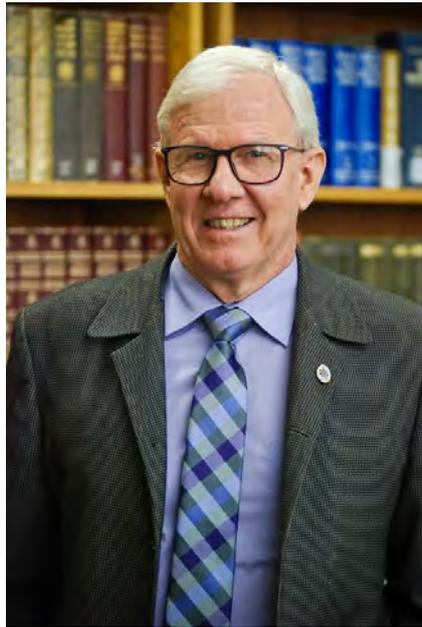
The full answers of both chatbots for prompt 6 are provided in the Addendum, available online (click [here](#)).

8. References

- Bothma, T.J.D. and I. Fourie.** 2024. Enhancing Conceptualisations of Information Behaviour Contexts through Insights from Research on e-Dictionaries and e-Lexicography. *Information Research. An International Electronic Journal* 29(2): 179-197.
<https://publicera.kb.se/ir/article/view/55127> (accessed 15 June 2025)
- Bothma, T.J.D. and I. Fourie.** 2025. Contextualised Dictionary Literacy, Information Literacy, and Information Behaviour in the e-Environment. *Library Management* 46(1-2): 14-28.
<https://doi.org/10.1108/LM-08-2023-0082>
- Bothma, T.J.D. and R.H. Gouws.** 2022. Information Needs and Contextualization in the Consultation Process of Dictionaries that are Linked to e-Texts. *Lexikos* 32(2): 53-81.
<https://doi.org/10.5788/32-2-1697>

- De Schryver, G.-M.** 2023. Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography* 36(4): 355-387.
<https://doi.org/10.1093/ijl/ecad021>
- Fuertes-Olivera, P.A.** 2024. Making Lexicography Sustainable: Using ChatGPT and Reusing Data for Lexicographic Purposes. *Lexikos* 34: 123-140.
<https://doi.org/10.5788/34-1-1883>
- Germain, T.** 2025. Is Google about to Destroy the Web? *BBC*, 13 June 2025.
<https://www.bbc.com/future/article/20250611-ai-mode-is-google-about-to-change-the-internet-forever> (accessed 15 June 2025)
- Gouws, R.H.** 2022. Extended Article Comments in Online Dictionaries. *Lexicography. Journal of ASIALEX* 9(2): 111-138.
<https://doi.org/10.1558/lexi.24289>
- Hartmann, R.R.K.** 1989. Sociology of the Dictionary User: Hypotheses and Empirical Studies. Hausmann, F.J. et al. (Eds.). 1989. *Wörterbücher. Dictionaries. Dictionnaires. An International Encyclopedia of Lexicography*: 102-111. Berlin: De Gruyter.
- Huete-García, Á. and S. Tarp.** 2024. Training an AI-based Writing Assistant for Spanish Learners: The Usefulness of Chatbots and the Indispensability of Human-assisted Intelligence. *Lexikos* 34: 21-40.
<https://doi.org/10.5788/34-1-1862>
- Lew, R.** 2024. Dictionaries and Lexicography in the AI Era. *Humanities and Social Sciences Communications* 11(1): 1-8.
<https://doi.org/10.1057/s41599-024-02889-7>
- Shneiderman, B.** 1996. The Eyes Have It: A Task by Data Type Taxonomy for Information Visualizations. *Proceedings 1996 IEEE Symposium on Visual Languages, Boulder, CO*: 336-343.
<https://doi.org/10.1109/VL.1996.545307>
- Wiegand, H.E.** 1977. Nachdenken über Wörterbücher. Aktuelle Probleme. Drosdowski, G., H. Henne and H.E. Wiegand (Eds.). 1977. *Nachdenken über Wörterbücher*: 51-102. Mannheim: Bibliographisches Institut.

Frederik Johannes Lombard 1953 – 2025



Dit was met groot leedwese dat die personeel van die Woordeboek van die Afrikaanse Taal (WAT) verneem het van die afsterwe van 'n geliefde kollega en vriend, dr. Frikkie Lombard, op 19 April 2025.

Dr. Frederik Johannes Lombard, of Frikkie soos ons hom almal geken het, het op 1 April 1982, ná drie jaar as onderwyser by Paul Roos Gimnasium, as mederedakteur by die WAT in diens getree. Hy het van vroeg af 'n passie vir taal, boeke en nuwe woorde en uitdrukkings gehad en 'n loopbaan in die leksikografie was nie 'n moeilike keuse nie.

Op 1 April 1993 is Frikkie tot senior mederedakteur bevorder. Dié pos het hy beklee tot hy op 1 Januarie 2001 in die pos van eindredakteur aangestel is. As daar 'n eienskap ten opsigte van woordeboekmaak is waaroor Frikkie by uitstek beskik het, was dit sy vlymskerp insig om betekenis akkuraat, helder en bondig te verwoord. Verder kon hy sinate moeiteloos knip en rangskik sodat hulle die definisie netjies ondersteun. Vanweë sy breë algemene kennis het die mederedakteurs hom met groot vrymoedigheid geraadpleeg — veral wanneer dit oor geskiedenis, weermagsake, fotografie, motor- en motorfietsterme, musiek of sport gehandel het — en almal is altyd met deernis en begrip behandel. Hy

het ook oor die jare as 'n mentor vir jonger kollegas opgetree en hulle altyd met die grootste geduld in die proses van manuskripmaak en -redigering begelei. Hy het boonop nie gekroom om erkentlikheid te bewys waar dit geregverdig was nie.

Vir 25 jaar, van 1995 tot 2020, het Frikkie in die Taalkommissie (TK) gedien, die eerste vyftien jaar as sekretaris en by geleentheid as waarnemende ondervoorsitter. As een van die langsdienende TK-lede het hy 'n betekenisvolle bydrae daartoe gelewer om die Afrikaanse spelreëls te help standaardiseer. Hiervoor is hy ook deur die Suid-Afrikaanse Akademie vir Wetenskap en Kuns met twee toekennings vereer.

Die onderwys het Frikkie ook altyd ná aan die hart gebly. Hy het sy doktorsgraad verwerf met die proefskrif *'n Metaleksikografiese fundering van Afrikaanse skoolwoordeboeke*. Frikkie het dan ook uiteindelik die geleentheid gekry om van hierdie teoretiese beginsels in die praktyk toe te pas deur mee te werk aan twee skoolwoordeboeke vir Oxford University Press, naamlik die eentalige *Oxford Afrikaanse Skoolwoordeboek* en die tweede uitgawe van die *Oxford Afrikaans-Engels English-Afrikaans Skoolwoordeboek School Dictionary*. Beide woordeboeke is kommersieel baie suksesvol en lewer steeds 'n belangrike bydrae om met die leesbegripkrisis in die Suid-Afrikaanse skoolstelsel te help.

Wat die werk aan die WAT betref, was Frikkie betrokke by tien van die sestien boekdele en ook die T-trajek wat slegs aanlyn bekendgestel is. Die laaste WAT-artikel wat Frikkie kort voor sy aftrede in Oktober 2023 vir redigering onder oë gehad het, was dié van die lemma **trek**, 'n polisemiese lemma met nie minder nie as 149 betekenisonderskeidings. In daardie stadium het ons nie gedink dat dit profeties sou wees nie. Vir sy aftrede het **trek** potensieel beteken "van huis, woning, kantoor of werkplek verander" en ook "'n tog wat met onderbrekings gepaard gaan, aanpak of onderneem". Dit was hom egter nie beskore nie. Uiteindelik het dit beteken dat hy finaal moes "versit of verskuif".

As mens buite die werkplek was sy liefde vir sy familie en vriende 'n inspirasie. Hy was baie mededeelsaam en sou altyd, wanneer hy iets as geskenk gekry het wat hy met ander kon deel, nie gekroom het om dit te doen nie. As grap- en storieverteller het Frikkie ook uitgeblink. Hy het 'n hele arsenaal grappies en anekdotes gehad waarmee hy vriende en kollegas vermaak het. Hy het 'n skerp humorsin en 'n ongelooflike geheue gehad en kon 'n storie op 'n baie interessante manier en in die fynste detail vertel.

Die personeel van die WAT het net die hoogste waardering vir die 42 jaar wat Frikkie Lombard die WAT met onderskeiding gedien het. Sy menslikheid, kollegialiteit en humorsin sal baie gemis word.

Alet Cloete
Woordeboek van die Afrikaanse Taal
Stellenbosch
Suid-Afrika
(aec@sun.ac.za)

<http://lexikos.journals.ac.za>; <https://doi.org/10.5788/35-2-2108> (Volle uitgawe / Full volume)
<http://lexikos.journals.ac.za>; <https://doi.org/10.5788/35-2-2095> (Lexicotribute)

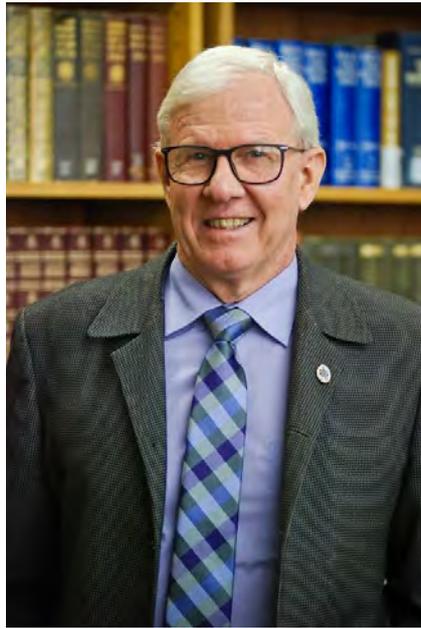
298 Frederik Johannes Lombard

Phillip Louw
Woordeboek van die Afrikaanse Taal
en
Departement Afrikaans en Nederlands, Universiteit Stellenbosch
Stellenbosch
Suid-Afrika
(pal@sun.ac.za)
(<https://orcid.org/0009-0007-2002-3081>)

en

Gerhard van Wyk
Woordeboek van die Afrikaanse Taal
Stellenbosch
Suid-Afrika
(gjvw@sun.ac.za)

Frederik Johannes Lombard 1953 – 2025



It was with great sadness that the staff of the *Woordeboek van die Afrikaanse Taal* (Dictionary of the Afrikaans Language, henceforth the WAT) learned of the passing of a beloved colleague and friend, Dr Frikkie Lombard, on the 19th of April 2025.

Dr Frederik Johannes Lombard, or Frikkie as we all knew him, joined the WAT as co-editor on the 1st of April 1982, after three years as a teacher at Paul Roos Gymnasium. From an early age, he had a passion for language, books and new words and expressions, and a career in lexicography was not a difficult decision.

Frikkie was promoted to senior co-editor on the 1st of April 1993. He held this position until he was appointed as final editor on 1 January 2001. If there was one quality in dictionary-making at which Frikkie excelled, it was his razor-sharp insight into expressing meaning accurately, clearly and concisely. Furthermore, he could effortlessly cut and arrange example sentences to neatly support the definition. Because of his broad general knowledge, co-editors consulted him with great confidence — especially concerning history, military affairs, photography, motoring and motorcycle terms, music or sport — and everyone

was always treated with compassion and understanding. Over the years, he also acted as a mentor to younger colleagues and guided them through the lexicographic process with the utmost patience. Moreover, he did not hesitate to show gratitude where it was warranted.

Frikkie served on the Taalkommissie (the Language Commission, hereafter TK) for 25 years, from 1995 to 2020, with the first fifteen of those years as secretary and occasionally as acting vice-chairman. As one of the longest-serving TK members, he made a significant contribution to the standardisation of Afrikaans spelling rules. For this, he was honored with two awards by the Suid-Afrikaanse Akademie vir Wetenskap en Kuns (South African Academy for Science and Arts).

Education also remained close to Frikkie's heart. He obtained his doctorate with the dissertation *'n Metaleksikografiese fundering van Afrikaanse skoolwoordeboeke* (A Metalexical Foundation of Afrikaans School Dictionaries). Frikkie later had the opportunity to apply some of these theoretical principles in practice by collaborating on two school dictionaries for Oxford University Press, namely the monolingual *Oxford Afrikaanse Skoolwoordeboek* (Oxford Afrikaans School Dictionary) and the second edition of the *Oxford Afrikaans-English English-Afrikaans School Dictionary*. Both dictionaries are commercial successes and continue to make an important contribution toward addressing the reading comprehension crisis in the South African school system.

As for the work on the WAT, Frikkie was involved in ten of the sixteen printed volumes as well as the stretch of the letter T, which was exclusively published online. The last WAT article edited by Frikkie shortly before his retirement in October 2023, was that of the lemma **trek**, a polysemic lemma with no less than 149 senses. At the time, we did not think that it would be prophetic. For his retirement, **trek** potentially meant "to change house, residence, office or workplace" as well as "to embark on or undertake a journey that involves interruptions". However, this was not meant to be his destiny. In the end, he had to "move or be moved" for good.

Outside of the workplace, his love for his family and friends was an inspiration. He was very generous and when he received something as a gift that he could share with others, he would always do so without hesitation. Frikkie also excelled as a joker and storyteller. He had an arsenal of jokes and anecdotes with which he entertained friends and colleagues. He had a sharp sense of humour and an incredible memory, and he could tell a story with remarkable detail and captivating flair.

The WAT's staff have only the highest appreciation for the 42 years that Frikkie Lombard served the WAT with distinction. His humanity, collegiality and sense of humour will be greatly missed.

Alet Cloete
Woordeboek van die Afrikaanse Taal
Stellenbosch
South Africa
(aec@sun.ac.za)

Phillip Louw
Woordeboek van die Afrikaanse Taal
and
Department of Afrikaans and Dutch, Stellenbosch University
Stellenbosch
South Africa
(pal@sun.ac.za)
(<https://orcid.org/0009-0007-2002-3081>)

and

Gerhard van Wyk
Woordeboek van die Afrikaanse Taal
Stellenbosch
South Africa
(gjvw@sun.ac.za)

Thematic Classification in Polish Electronic Dictionaries — From Tradition to New Tools and Opportunities*

Barbara Batko-Tokarz, *Philological Faculty,
Jagiellonian University, Krakow, Poland*
(barbara.batko-tokarz@uj.edu.pl)
(<https://orcid.org/0000-0001-9147-5180>)

Abstract: The main topic of this article is the thematic classification of vocabulary in two Polish dictionaries that were compiled at the beginning of the 21st century. The first is the *Wielki słownik języka polskiego PAN* (WSJP PAN) [The Polish Academy of Sciences Great Dictionary of Polish], which describes contemporary vocabulary and can function as a thematic dictionary. The second is the *Słownik pojęciowy języka staropolskiego* (SPJS) [Conceptual Dictionary of Old Polish], which collects vocabulary up to the end of the 15th century and arranges it in conceptual fields. The primarily electronic format of these dictionaries has created completely new possibilities for classification according to meaning, which is worth briefly describing in the context of the thematic lexicography used in Polish paper dictionaries to date. Therefore, the most important Polish thematic dictionaries over the centuries are listed first (Section 2), followed by a more detailed description of the SPJS and WSJP PAN (Section 3). In Section 4, the following is described based on these two Polish electronic dictionaries: the ease of switching between onomasiological and semasiological orders, the significantly greater possibilities for classifying lexical units, and the ability to search fields using advanced search functions. These considerations are important given the growing number of electronic dictionaries increasingly adopting onomasiological orders.

Keywords: ONOMASIOLOGICAL LEXICOGRAPHY, THEMATIC LEXICOGRAPHY, THEMATIC DICTIONARIES, CONCEPT DICTIONARIES, PRIMARILY ELECTRONIC DICTIONARIES, SEARCH TOOLS IN ELECTRONIC DICTIONARIES, ONOMASIOLOGICAL AND SEMASIOLOGICAL ORDER, THEMATIC CLASSIFICATION OF VOCABULARY

Zusammenfassung: Thematische Klassifikation (des Wortschatzes) in polnischen elektronischen Wörterbüchern — von der Tradition bis hin zu neuen Werkzeugen und Möglichkeiten. Der vorliegende Artikel befasst sich mit der thematischen Klassifizierung des Wortschatzes in zwei zu Beginn des 21. Jahrhunderts erarbeiteten Wörterbüchern

* This publication and its translation has been supported by a grant from the Priority Research Area (POB Heritage) under the Strategic Programme Excellence Initiative at Jagiellonian University  .

der polnischen Sprache. Das erste ist *Wielki słownik języka polskiego PAN* (WSJP PAN) [Großes Wörterbuch der polnischen Sprache der Polnischen Akademie der Wissenschaften], das den zeitgenössischen Wortschatz ab dem Jahr 1945 erfasst und zugleich als thematisches Wörterbuch genutzt werden kann. Das zweite, *Słownik pojęciowy języka staropolskiego* (SPJS) [Begriffswörterbuch des Altpolnischen], dokumentiert den bis zum Ende des 15. Jahrhunderts verwendeten Wortschatz, geordnet nach Begriffsfeldern. Die primär elektronische Form beider Wörterbücher eröffnet völlig neue Möglichkeiten für die Klassifikation nach Bedeutungen. Diese Entwicklungen verdienen eine kurze Darstellung im Kontext der thematischen Lexikografie, wie sie bislang in gedruckten polnischen Wörterbüchern praktiziert wurde. Daher werden zunächst im zweiten Abschnitt die wichtigsten thematischen Wörterbücher des Polnischen im Verlauf der Jahrhunderte vorgestellt. Im dritten Abschnitt folgt eine detaillierte Charakterisierung des SPJS und des WSJP von PAN. Der zentrale vierte Abschnitt des Artikels widmet sich den verschiedenen Zugriffsformen auf das Wörterbuchmaterial in diesen beiden polnischen elektronischen Wörterbüchern. Besonders hervorgehoben wird dabei die Möglichkeit des einfachen Wechsels zwischen onomasiologischer und semasiologischer Ordnung. Beschrieben werden zudem deutlich erweiterte Optionen für die Klassifizierung des sprachlichen Materials sowie für die feldspezifische Suche mithilfe erweiterter Suchfunktionen. Diese Überlegungen gewinnen an Relevanz angesichts der wachsenden Zahl elektronischer Wörterbücher, die zunehmend auf einer onomasiologischen Ordnung basieren.

Schlüsselbegriffe: ONOMASIOLOGISCHE LEXIKOGRAFIE, THEMATISCHE LEXIKOGRAFIE, THEMATISCHE WÖRTERBÜCHER, BEGRIFFSWÖRTERBÜCHER, PRIMÄR ELEKTRONISCHE WÖRTERBÜCHER, SUCHFUNKTIONEN IN ELEKTRONISCHEN WÖRTERBÜCHERN, ONOMASIOLOGISCHE UND SEMASIOLOGISCHE ORDNUNG, THEMATISCHE KLASSIFIKATION DES WORTSCHATZES

1. Introduction

The aim of this article is to present the thematic classification of vocabulary used in two Polish dictionaries that were compiled at the beginning of the 21st century (and are only available online), as well as new possibilities for its use as offered by their electronic versions. The first of these is the *Wielki słownik języka polskiego PAN* (WSJP PAN) [The Polish Academy of Sciences Great Dictionary of Polish], which describes contemporary vocabulary from 1945 onward. The second is the *Słownik pojęciowy języka staropolskiego* (SPJS) [Conceptual Dictionary of Old Polish], which collects vocabulary up to the end of the 15th century. These dictionaries were created independently of each other, as they were developed by two different teams of lexicographers, and they differ in the types and amounts of linguistic material covered. However, they also have much in common, in that they are primarily electronic dictionaries that use both semasiological and onomasiological orders. Dictionary entries are also available thematically, as both dictionaries employ semantic classification of large, collected vocabulary groups. Due to the pioneering nature of these solutions, the authors responsible for this kind of classification in each dictionary have presented both the methodological framework and practical aspects of applying these classifications, situating

them within the broader context of Polish and European thematic or onomasiological lexicography (Batko-Tokarz 2019; Sieradzka-Baziur 2020).

Dictionaries that do not follow alphabetical but semantic order constitute a subtype of onomasiological lexicons (Hartmann and James 2002: 101-102). In European lexicography (including Slavic), the following terms are used to describe such works: *analogical*, *conceptual*, *ideographical*, *ideological*, *thematic*, *topical*, *systematic*, *semantic (dictionary)*, *nomenclator*, and *thesaurus* (e.g., Jedlička 1977: 222-223; Sierra 2000: 224; Nikolić-Hoyt 2004: 7-8, 21; Hüllen 2006: 13; França 2010: 1291-1292; Obstová, Tichý and Klégr 2022: 77).¹ There are also differences in the terminology used in both dictionaries, as the WSJP PAN uses the terms *thematic classification* or *thematic qualification*, *thematic dictionary*, *thematic zones*, *fields*, and *subfields*, while the SPJS uses the term *conceptual* in its title, *conceptual categories* in its classification scheme, *semantic unit* (Sieradzka-Baziur 2020: 46) and *semantics* in entries.

In this article, terms containing the adjective *thematic* — such as *thematic classification*, *thematic dictionary*, and *thematic lexicography* — are used as overarching in relation to other types of dictionaries and classifications that arrange vocabulary according to its semantic properties. This is firstly related to the tradition of such terminology in Polish and Slavic lexicography (although other terms are also used in this tradition).² Additionally, this decision is compliant with new lexicographic trends because, as analyses of the interest in onomasiological dictionaries at the turn of the 20th and 21st centuries indicate, the adjective *thematic* is becoming increasingly more popular due to its general meaning (Kawalya and De Schryver 2013: 178). These terms will be understood broadly (McArthur 1998; Hartmann and James 2002: 142; Burkhanov 2010: 242-243 — the first meaning);³ that is, they refer both to dictionaries with a simple structure (such as those arranging words according to specific topics important in teaching, or presenting a certain type of vocabulary) and to general language dictionaries with precisely built hierarchical structure (such as thesauruses or conceptual lexicons, as their subtypes).

SPJS and WSJP PAN are available only in electronic format, which firstly means that it is easy to switch from the order typical of semasiological dictionaries to that used in onomasiological dictionaries, and secondly, advanced search options are available. It is worth showing, using the example of these two Polish dictionaries, how the use of electronic tools makes the thematic classification used in dictionaries much more effective and functional. The electronic form of dictionaries enables a completely new way of functioning for Polish thematic lexicography, which until now has only existed in printed form. Furthermore, it eliminates many of its previous limitations and difficulties, which is why more attention to the connections between lexicography and new disciplines, especially information technology, is necessary (Hartmann 2005: 15-16).

However, it would be difficult to demonstrate this without a brief background overview. This overview includes a short description of the tradition of Polish thematic lexicography over the centuries, with particular emphasis on Polish printed dictionaries of the 20th and 21st centuries, arranged in whole or

in part thematically in line with the words: "Yet, it seems high time that non-alphabetical dictionaries should at least regain their history" (Hüllen 2006: 21). Although the tradition of lexicographic works arranged in such a way may differ depending on language, the cross-sectional approach to a specific type of dictionaries in a given country is justified. Such an approach makes it possible to identify what is universal, not only in terms of historical practices, but also in relation to future applications. Therefore, the most important Polish printed thematic dictionaries in history will be briefly presented below (Section 2), followed by a description of contemporary electronic (online) dictionaries using a thematic layout, namely the SPJS and WSJP PAN (Section 3). In the penultimate and particularly important part of this article using these two dictionaries as examples, new possibilities and tools for the thematic classification of vocabulary offered by electronic lexicography will be presented (Section 4).

2. Brief history of Polish thematic lexicography⁴

2.1 From the 16th century to the first half of the 20th century

The tradition of arranging vocabulary in dictionaries according to meaning is older than the European culture of writing — dictionaries arranged thematically rather than alphabetically constitute the foundations of European lexicography (cf. Nikolić-Hoyt 2004: 39-62; Hüllen 2006: 32-54). The beginnings of Polish lexicography are also associated with the thematic arrangement of vocabulary, which can be seen in the first printed dictionaries (with Polish materials modeled on foreign works). This tradition began with the Polish adaptation of the dictionary by Dutch philologist Johannes Murmellius (Hüllen 2006: 324-326), namely the German–Latin–Polish dictionary titled *Dictionarius Ioannis Murmelli variarum rerum* (Murmeliusz 1526), which was frequently reprinted back then. It is now believed that the Polish part was prepared by Hieronim of Wieluń (Spiczyński), but in the Polish tradition this work is referred to as the "Murmellius Dictionary". In 1528, Francis Mymer's *Dictionarium trium linguarum* was published with the same language combination (Mymer 1528). The author modeled it on a Czech dictionary from 1513 and, in part, on Murmellius' dictionary (Gruszczyński and Saloni 2013: 207-208). This type of vocabulary arrangement was also used in the so-called *Słownik* (1532), an adaptation of the quadrilingual dictionary (Gruszczyński and Saloni 2013: 208). The last thematically organised dictionary of the 16th century was the Latin–German–Polish *Nomenclator selectissimas rerum appellationes* [...] by Petrus Artomius (Artomiusz 1591), which was published as a reworking of the Latin–German version. This became the most frequently reprinted Polish dictionary in the 17th century, though most of the editions were erroneously signed with the name of Murmellius (Gruszczyński and Saloni 2013: 209).

From the 17th century onward, the alphabetical arrangement began to dominate European lexicography, including that of Poland (Gruszczyński and Saloni 2013: 209 ff.). Thematically arranged dictionaries from the 17th to the mid-20th

centuries were rare and considered anachronistic by the lexicographic standards of the time — most often, these were glottodidactic dictionaries. However, it was then that a work of great importance for Polish thematic lexicography was created, namely, the *Dobór wyrazów. Słownik wyrazów bliskoznacznych i jednoznacznych. Do praktycznego użytku* [Selection of Words. Dictionary of Synonymous and Mono-semantic Words. For Practical Use] (Zawiliński 1926). The author based his work on the 1912 edition of Roget's *Thesaurus*. However, this dictionary was underestimated by its first users and researchers, who were unprepared for its reception and identified it merely as a dictionary of synonyms, despite its much higher quality.

2.2 From the second half of the 20th century to modern times (printed dictionaries)

Since the middle of the 20th century, there has been a revival of the tradition of classifying vocabulary based on meaning, alongside its increasing popularity in linguistic research and lexicography. Furthermore, the influence of foreign achievements in this field on Polish lexicography is evident, as is the development within the country itself. The paper dictionaries created at that time using thematic arrangements can be divided into three types.

First, they were local dialect dictionaries. In the 1960s, the *Porównawczy słownik trzech wsi małopolskich* [Comparative Dictionary of Three Lesser Poland Villages] was published (Kucała 1957). The researcher followed the scheme of R. Hallig and W. von Wartburg (1952) but adapted it to the collected vocabulary. This dictionary gained recognition in Polish and Slavic dialectology. Although the lexicography of dialects departed from thematic arrangements in later years, there has recently been a return to this method of ordering the dictionary macrostructure. In this respect, at least two large dictionary projects are worth mentioning, with the first being the ongoing ethnolinguistic *Słownik stereotypów i symboli ludowych* [Dictionary of Folk Stereotypes and Symbols] (Bartmiński and Niebrzegowska-Bartmińska 1996–2024). According to the plan, the work was to consist of seven overarching subject areas, starting with nature and ending with man, culture, and categories that organise the world. Another dictionary of this type is the recently completed 12-volume *Słownik gwar Lubelszczyzny* [Dictionary of Lublin Dialects] (Pelcowa 2012–2024). The overarching thematic arrangement of the individual volumes was intended to reflect the image of the world and man as preserved in the consciousness of the rural population. At present, increased attention is being devoted to electronic versions of dictionaries in Polish dialect lexicography (cf. Rembiszewska 2016: 49, 80-81), as well as to the challenges and advantages associated with employing thematic organisations within these resources (cf. Rembiszewska 2016: 42, 223-225, 273). However, to date, no dictionary has yet combined thematic classification with an electronic format.

Second, thematic arrangements were used in the dictionaries of the different language varieties in Polish lexicography in the second half of the 20th century, for example the *Słownik gwary studenckiej* [Lexicon of Students' Dialect] (Kaczmarek, Skubalanka and Grabias 1994),⁵ and the *Słownik polszczyzny potocznej* [Colloquial Polish Dictionary] (Anusiewicz and Skawiński 1996). Because the entry orders in these dictionaries were based on hierarchical classification schemes created from the bottom up, they showed the image of non-linguistic reality and the value systems that were encoded in these language varieties. At the beginning of the 21st century, the use of thematic organisation in Polish lexicography in phraseological dictionaries and dictionaries of proverbs became prominent, which made it easier to show the valuations that were encoded in these language varieties, for example the *Wielki słownik frazeologiczny* [Great Phraseological Dictionary] (Lebda 2009).

Third, pedagogical dictionaries are an important branch of thematic lexicography in Poland (as is the case in world lexicography) (Burkhanov 1999; Hartmann and James 2002: 107), even though many of them are arranged alphabetically. Bilingual thematic dictionaries that include Polish have been very popular in Poland since the early 1990s. However, these types of dictionaries are not as popular in other Slavic countries, such as Serbia and Croatia, as they are in Poland, since thematic arrangements are more commonly found in the form of picture dictionaries in these countries (cf. Daković 2011: 40-45). In Polish lexicography, however, these dictionaries are rare, although an original and highly rated glottodidactic publication of this type does exist, titled *A co to takiego? Obrazkowy słownik języka polskiego* [What's This? A Picture Dictionary of Polish] (Seretny 2008), which was first published in 1993. It is also worth mentioning monolingual pedagogical dictionaries that contain basic vocabulary for learners of Polish, which are organised thematically, for example the *Ilustrowany słownik podstawowy języka polskiego* [Illustrated Elementary Dictionary of Polish] (Kurzowa 2020),⁶ first published in 1999. Monolingual pedagogical dictionaries intended primarily for native speakers interested in improving their language skills should also be mentioned, such as the *Słownik tematyczny języka polskiego* [Thematic Dictionary of Polish] (Kita and Polański 2002).

There remains a need for electronic thematic dictionaries (monolingual, bilingual, or multilingual) of a glottodidactic nature, despite the existence of online thematic dictionaries that are added to language portals or multilingual thematic portals that are created by users (Daković 2011: 44-45). However, printed formats still dominate among dictionaries published by established publishing houses, and their electronic versions are rarely made available.⁷ Linguists working in Polish scientific centres have been developing various ideas and projects related to bilingual (Daković 2011: 45-47) and multilingual corpora-assisted thematic dictionaries (Koutny 2005). Researchers are also paying more attention to both the theoretical and practical potentials of pedagogical thematic dictionaries regarding the enhancement of the linguistic or cultural skills of their users (Koutny 2012; Kononenko and Mytnik 2017; Sękowska 2020; Greń 2023).

3. Newest electronic dictionaries in which thematic classification of vocabulary is applied

In a survey on onomasiological lexicons in 20th-century Europe, Hartmann (2005: 11) indicated the potential of two particular Slavic languages in relation to onomasiological lexicography: Russian and Polish, which achieved the best results among the Slavic languages. As far as Russian is concerned, this lexicography is particularly related to its very large user base, but it is also related to the academic tradition of creating such dictionaries (Hartmann 2005: 11). In the case of Polish (which has a significantly smaller number of users), Hartmann (2005: 13) clearly emphasised the potential of Polish lexicographers. He wrote the following in the paragraph that was dedicated to Polish onomasiological lexicons: "Compilers and theoreticians have been open to ideas and innovations from both East and West, a development that is likely to continue into the future" (Hartmann 2005: 13). These predictions turned out to be correct, because work began on two Polish electronic dictionaries soon after the publication of this article, and they employed the thematic classification of vocabulary that is described in this article.

3.1 The *Słownik pojęciowy języka staropolskiego* (SPJS) [Conceptual Dictionary of Old Polish]

In 2015, a project was completed at the Institute of Polish Language of the Polish Academy of Sciences in Krakow (which began in 2011). Its outcome was the creation of the SPJS. It is an electronic dictionary that includes the lexis of the earliest period of Polish language development up until the end of the 15th century. Based on the printed dictionary of old Polish language in alphabetical order, the first comprehensive and systematic lexicographic presentation of old Polish semantics was created (Sieradzka-Baziur 2020: 9). As emphasised by the dictionary's editor-in-chief, Bożena Sieradzka-Baziur (2020: 41), the SPJS is a "digitally born dictionary" because "it retains its distinctiveness from the *Old Polish Dictionary* — it has a different structure, and the material taken from that work has undergone extensive modifications, which include both the macrostructure of this dictionary and the microstructure of the entries".

In the SPJS, more than 23 000 medieval words and several thousand expressions were classified, but the classifications (sometimes even to several categories) were made on the level of meanings, which often resulted in several distinct classification within a single dictionary entry (Sieradzka-Baziur 2020: 49, 72). For practical reasons, the lexicographers decided to present the old Polish linguistic universe within the field of contemporary concepts. Again in Polish thematic lexicography (Kucala 1957), the scheme that was applied in the dictionary by R. Hallig and W. von Wartburg (1952) became the starting point for creating a classification structure in the SPJS. However, the original scheme was modified and adjusted to old Polish vocabulary, so the differences between them were clear (Sieradzka-Baziur 2020: 69-75). The classification structure consisted of more than 200 con-

ceptual categories within the hierarchised classification scheme. Similarly to the original dictionary, this comprised of three general areas: *Wszechświat i świat* [Universe and the World]; *Człowiek* [Human], and *Człowiek a wszechświat i świat* [Man, the Universe and the World]. Although the work on the SPJS is finished, the lexicographers are open to its possible modifications (Sieradzka-Baziur 2020: 83).

In the bookmark named *Słownik* [Dictionary], users can search for words by using the conceptual classification scheme (which is the essence of the project). However, this can also be done according to an alphabetical arrangement or grammar characteristics. A general search window is also available in the dictionary. The dictionary's website has recently made the English version of the classification scheme available, including information on the number of semantic units assigned to each category (<https://spjs.ijppan.pl/spjs/strona/classificationScheme>).

3.2 The *Wielki słownik języka polskiego Polskiej Akademii Nauk (WSJP PAN)* [Polish Academy of Sciences Great Dictionary of Polish]

The WSJP PAN was created in 2007 as the first primarily electronic general dictionary of contemporary Polish. This lexicographical project is still under development. Its history and principles of preparation are also described in English (Żmigrodzki 2018).

In the description, modern solutions implemented in the project are outlined, including the thematic classification introduced into newly created dictionary entries. In July 2024, the total number of all the entries in the dictionary was more than 90 000. Meanwhile, work has also been completed on a dictionary application for mobile devices, which can be downloaded and used on a tablet or mobile phone (cf. <https://wsjp.pl/page/historia>). The vast majority of the autosemantic entries (i.e., those in which semantic classification was possible), received thematic categorisation, although sometimes they could be assigned to the *athematic* category. Thematic classification was also carried out on the level of particular meanings that were separated within the dictionary entries, and it was possible to classify them into several thematic categories (Batko-Tokarz 2008: 34-39; 2018: 34-36).

For the purpose of the thematic classification of the entries that are being created in the WSJP PAN, an original classification scheme was created that is hierarchical and consists of three degrees. This comprises seven general thematic groups: (1) Man as a Physical Being; (2) Man as a Psychological Being; (3) Everyday Life of a Man; (4) Man in Society; (5) Man and Technology; (6) Man and Nature; and (7) Physical Categories. Within these groups, 45 fields are separated, followed by 205 subfields, which are the basic place of the lexical units classification (although some units may be classified on higher levels). The thematic classification was intended to cover newly created dictionary entries, therefore the classification scheme had to be designed *a priori*. To improve the functionality of such a classification, the names of certain categories were revised, and their scope was made clearer and more understandable for users in 2013.

The rules of the classification and the classification scheme have also been described in detail (Batko-Tokarz 2008; 2019: 135-293).

The dictionary makes it possible to access the lexicographic material in different ways. One can check words according to an alphabetical arrangement, but there are also various advanced search options, with the possibility of marking various categories (including thematic ones). Thanks to this solution, it was possible to find dictionary entries that were classified into particular groups and related fields as well as thematic subfields for many years (Żmigrodzki 2018: 215-216; Żmigrodzki 2022: 157-160). At the end of 2022, the bookmark *Słownik tematyczny* [Thematic dictionary] was added to the top right corner of the home page, which shows how important such access to the vocabulary was for the lexicographers. The rules for using the dictionary are clearly presented on the website in the *Jak korzystać* [How to use] section, which also includes instructional videos demonstrating how to use the various tools available to users. They were created as part of a modern online guide to this dictionary. One of them focuses to the thematic dictionary (<https://www.youtube.com/watch?v=0tySvF1ZKBA>)

4. Polish thematic lexicography in the electronic form — New tools and opportunities

The advantages and disadvantages of primarily electronic dictionaries are increasingly more interesting to various researchers (cf. Granger and Paquot 2012; Ferrett and Dollinger 2021) — including those that deal with Slavic languages (cf. Żmigrodzki 2008: 98-123; Balalaieva 2021). The electronic format results in there being no limits to volume as well as the linear arrangement of dictionaries no longer being necessary. Instead, lexicographers are dealing with multimedia data-bases and open, flexible, and interactive systems with various search tools. User may adjust systems' structures to their needs by opening the desired modules. Moreover, electronic dictionaries (especially those that are included on the internet) are available to everyone, and their forms make it possible for lexicographers to implement the necessary modifications and enhancements. Thematic classification therefore seems to be much more effective in an electronic dictionary than in a paper one. Considering this, the advantages and new opportunities of this combination are presented below on the base of the WSJP PAN and SPJS.

4.1 Combining a semasiological dictionary with an onomasiological dictionary

In an electronic dictionary, there are no longer differences among various dictionaries that used to function according to the principle of opposition. In modern printed dictionaries, the divisions into semasiological, onomasiological, and picture dictionaries have also become increasingly more blurred (Kawalya and De Schryver 2013: 180, 182). In electronic lexicography this is much simpler,

which is why Kawalya and De Schryver (2013: 185-186) suggest a new useful model of a multilingual dictionary. This model is called the "alphaconceptual+dictionary", which combines the advantages of these three types of dictionaries and eliminates the faults of each of them separately (cf. Kawalya and De Schryver 2013: 185-186).

In both of the aforementioned Polish electronic dictionaries, the lexicographers managed to combine the arrangement that is typical of thematic dictionaries with information that can be found in a semasiological dictionary,⁸ and moving from a single entry to a thematic category (and the other way around) is very easy. In both of these dictionaries, the structure of an entry makes it possible to check into which fields of the classification scheme a given unit was classified. In the WSJP PAN the arrangement within an entry, or within its individual meaning (subentry), is structurally organised in the form of expandable tabs. The *Kwalifikacja tematyczna* [Thematic qualification] tab appears below the *Definicja* [Definition] tab (cf. Żmigrodzki 2018: 216; Żmigrodzki 2022: 157-158). When opened, it displays information about the thematic classification (Figure 1).

The screenshot shows the website interface for the 'Wielki słownik języka polskiego' (WSJP PAN). The main content area displays the entry for 'dom' (house). Below the main definition, there is a subentry '1. budynek' (building). A flowchart illustrates the thematic classification of 'budynek':

- CODZIENNE ŻYCIE CZŁOWIEKA
- ↓
- Najbliższe środowisko życia człowieka
- ↓
- dom/inne miejsca zamieszkania i ich otoczenie
- ↓
- CZŁOWIEK I TECHNIKA
- ↓
- Budownictwo
- ↓
- rodzaje budowli i ich cechy

On the right side, a vertical menu of tabs is visible, with 'Kwalifikacja tematyczna' (Thematic qualification) selected and highlighted in green. Other tabs include 'Definicja', 'Połączenia', 'Cytaty', 'Odmiana', and 'Pochodzenie'.

Figure 1: The *Kwalifikacja tematyczna* [Thematic qualification] tab in entry *dom* [house] (1st subentry 'building') in the WSJP PAN

In the SPJS, all relevant information within an entry (subentry) is displayed on the right-hand side, including classification into a conceptual category in the *Semantyka* [Semantics] tab (Figure 2).

Słownik pojęciowy języka staropolskiego

O słowniku Słownik Wyszukiwarka

artykuł hasłowy: dom

Jednostki:	Szczegóły:
(1) 1. <i>budynek mieszkalny (często wraz z mieszkańcami i przynależnym gospodarstwem), samo mieszkanie, miejsce stałego pobytu, aedificium, domus (saepe una cum habitatoribus et re familiar), habitanti locus stabilis, cubiculum, interna domus spatia</i>	Typ Jednostka pochodzi ze Słownika staropolskiego
(2) –	Rodzaj autosemantyczna
(3) – 'kryjówka'	Numer (1) 1.
(4) – <i>dubium</i>	Definicja 'budynek mieszkalny (często wraz z mieszkańcami i przynależnym gospodarstwem), samo mieszkanie, miejsce stałego pobytu, aedificium, domus (saepe una cum habitatoribus et re familiar), habitanti locus stabilis, cubiculum, interna domus spatia'
(5) 2. <i>'świętynia, templum'</i>	Gramatyka rzeczownik
(6) –	Semantyka 2.4.6.4. BUDOWNICTWO. ZAMIESZKIWANIE 2.4.6.4.1. BUDOWLA. BUDYNEK. DOM etc.
(7) –	Przykład w transliteracji Orądował 1śsm dę gich dōmū, by penądze brali, a ony gich nechczeli wżącz 1395
(8) –	Przykład w transkrypcji Orędował j<e>śm do jich dōmu, by pieniądze brali, a oni jich nie chcieli wziąć 1395
(9) –	Lokalizacja <i>Leksy II nr 1773</i>
(10) –	Uwagi [Cytat został uzupełniony o fragment a ony gich nechczeli wżącz.]
(11) –	
(12) 3. <i>'zakład, instytucja, aedificia, vario publicoque apta usu': dom kupiecki 'sklep, bazar, taberna mercatoris'</i>	
(13) – <i>'skład, sklep z lekami, korzeniami i kosmetykami, apotheca, emporium'</i>	
(14) – <i>'dom noclegowy, hospitium'</i>	
(15) – <i>'domus conductitia'</i>	
(16) – <i>'dom publiczny, lupanar'</i>	
(17) 4. <i>vód, rodzina, gens, familia'</i>	

Figure 2: Entry *dom* in the SPJS (1st subentry 'building')

From the side of single entries, such a classification scheme facilitates the semantic characteristics of a given word, and it may emphasise its polysemousness when particular meanings obtain different classifications (partially or fully), which is often the case. Such information is valuable for people who are learning a new language, but it is also good for other users (Batko-Tokarz 2018: 82-83).⁹ This was especially important in the SPJS due to the nature of the vocabulary that was collected in the dictionary — and assignments to particular fields is a valuable supplementation of a definition. The dictionary user also has access to definitions and other elements of the microstructure, for example in the WSJP PAN an entry includes large collections of various semantic, grammatical, and etymological information, as well as collocations and full sentence quotations (Koziół-Chrzanowska 2017: 204; Żmigrodzki 2018: 211). In the SPJS, grammatical characteristics as well as examples of word use in medieval texts are provided (Sieradzka-Baziur 2020: 59-63).

If dictionary users are interested in a given meaning, they can move to the field into which it was classified to see the units that are thematically related to this topic. In both dictionaries, users can search for entries according to a hierarchical thematic index that gives users access to the linguistic material in the fields, even though it looks slightly different in terms of the graphics in each dictionary. In the SPJS, the whole long and complex structure becomes visible in the bookmark *Schemat klasyfikacyjny SPJS* [SPJS Classification scheme]. Within the *Słownik tematyczny* [Thematic dictionary] section of the WSJP PAN, users

are presented with an expandable hierarchical structure of thematic scheme (Figure 3). Clicking the question mark icon on this page provides instructions on how to use the search engine and the thematic classification principles used in this dictionary.



Figure 3: The *Słownik tematyczny* [Thematic dictionary] in the WSJP PAN

If users search for words using the classification scheme, they can see lexical units assigned to various thematic areas, fields and especially subfields (in alphabetical arrangement). After choosing the unit, they can quickly move to a description that is typical of a semasiological dictionary (Figures 4 and 5).

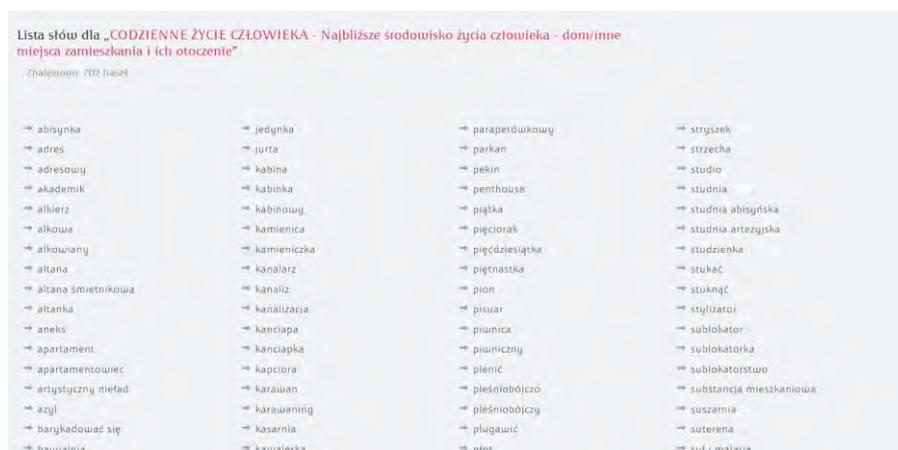


Figure 4: Units classified into the subfield *Dom/inne miejsca zamieszkania i ich otoczenie* [Home/other places of residence and their surroundings] in the WSJP PAN (fragment view; the full list contains 702 entries)

2.4.6.4.1. BUDOWLA. BUDYNEK. DOM etc. - jednostki:

baszta (1); bóżnica (1); buda (1); buda (2); budowanie (2);
cerkiew, cerki, cyrkiew (1); cerkiew, cerki, cyrkiew (2);
chalupa (1); chlew (6); chycz, częściej pl. tantum chycze (1);
chyczec a. chyczek (1); chyż, chyża (1); chyżek (1); chyżka
(1); chyżka (2); ciemnica (1); ciemniczny (1); ciemność (3);
cieplica, cieplice (2); ciosanie (2); ćma (4); dom (1); dom (3);
dom (5); dom (12); dom (14); dom (15); dom (16); doma (1);
domek (1); domek (2); (S) domostwo, domowstwo (1);
domowny (1); domowy (4); dworowy (1); dworzysko (3);
dwór (1); dwór (3); dwór (5); dwór (7); dymowe (1); folwark,
folwerk, forwak, forwalk, forwark, wolwark (1); folwarkowy,
forwarkowy (1); gałka (2); glinianka (1); gmach (1); gnojny
(1); gołębiniec (1); gospoda (1); gospoda (2); gospodarz (5);
gospodarz (6); gospodarzyć (1); gospodka (1); gospodyni
(2); gospodzić (1); gościniec (3); gościnny, gościni, gościny
(2); gość (3); grodziec (1); gród (1); izba, histba, istba, izdba
(1); jata (1); kamienica (1); kaplica (1); karczernikow(y) (1);
karczemny (1); karczma, kaczma (1); karczmisko (1); kąt (3);
kiczka (2); kikić (?) (1); klasztor (1); klatka, kietka (1); klec
(1); kluza (1); kmiećnica (?) (1); komnata, kownata (1);
komora (1); komora (5); komórka (1); konica (1); kościół (1);
kościół (4); kościółek, kościółek (1); kram, krom (1);
kramnica (1); kramny (1); kramny (3); kruchta, krukta (1);
kruchta, krukta (2); kruchta, krukta (3); kryłos (1); kuczka
(1); 1. kupiecki, kupiecki (1); kupiectwo (1); kupny (6);
łakwa (1); lekarski (1); lonar (1); ława (2); łęczysko (2);
łęczyszczce (2); modlebnia (1); modlebnica (1); modlitwa,
moglitwa (4); mrowanie (1); mrowany (1); namiot (1);
nasuć (4); nieczysty (5); niepoczesny, niepoczesny (2); noc
(3); noc (4); noc (5); nocleg (1); nocować (1); obcowanie,
hobcowanie (5); obora (1); obręba (2); odpoczynienie,
otpoczynienie (4); odpoczywać, odpoczywać (2); okół (2);
oratarz (1); pacisz (?) (1); pałac (1); pałacowy (1); piekuta (1);
plac, plec (1); płot (6); pobudowanie (1); podniebie (2);

Figure 5: Units classified into the category *Budowla. Budynek. Dom* etc. [Structure. Building. House etc.] in the SPJS (fragment view; the full list contains 249 entries)

The electronic form of a dictionary helps lexicographers overcome the faults of the thematic order that are visible in the aforementioned printed Polish thematic dictionaries, and this seems to be the solution to problems of metaideography (Burkhanov 1999: 144-147). Such faults may include a frequent lack of definitions of thematically classified units, which results in certain doubts that are related to the reasons for their classification. When classification in printed dictionaries was carried out at the level of individual meanings, the various senses of a single word were distributed across different sections of the scheme. Problems also include the arbitrariness and non-intuitiveness of a given classification or its assignments — especially when there are no definitions — as well as related difficulties when searching for particular units by the thematic scheme. Researchers have mentioned that those units that are thematically non-classifiable or hard to classify have been omitted in such dictionaries (cf. Batko-Tokarz 2019: 114-122). Until now, the lexicographers have combined various kinds of access to the dictionary material through thematic or alphabetical indices in order to remove the faults, and they have used the advantages of both arrangements of a dictionary macrostructure (Burkhanov 2010: 108-109). This was dealt with in the above-mentioned Polish dictionaries (including the historical ones), but they have not solved all of the problems (Batko-Tokarz 2019: 122-124).

4.2 Abundant and rich linguistic material in thematic fields and various opportunities for its characteristics

Both of the above-mentioned Polish dictionaries follow the assumptions that thematic classification should be made on the level of meanings separated within an entry, and that there can be several possible classifications. Due to these assumptions, thematic classification was able to include a considerable number of lexical units. First, in the dictionaries of general Polish, entry articles often have several or even more than a dozen meanings. Second, many of them have been classified several times. According to the recent official statistics of the WSJP PAN¹⁰ of as many as 38 215 units were simultaneously classified into two fields, and 4 001 units were classified into three or more fields (Żmigrodzki 2022: 156).

The electronic form of a dictionary definitely facilitates broadly understood thematic classification, as there are no limitations to the volume of such a lexicon. Because of this, the WSJP PAN is the largest thematic dictionary of contemporary Polish at the moment and outrivals the other Polish thematic dictionaries mentioned above. According to the status as of August 2021, it included 141 561 meanings that were classified in a thematic manner. If multiple classifications are, however, taken into account, it turns out that the total number of classifications to thematic categories is as many as 188 037 (Żmigrodzki 2022: 155, 160). Although it refers to the early stage of the development of Polish, the SPJS also includes 80 594 thematic classifications, while 44 177 units were classified thematically into at least one (and often more) categories (Sieradzka-Baziur 2020: 90). A similar number of units are included in the largest translation dictionaries that are arranged thematically and include Polish, while the above-mentioned contemporary Polish glottodidactical thematic dictionaries (Kita and Polański 2002; Kurzowa 2020) or phraseological dictionaries (Lebda 2009) count from several to more than ten thousand entries (Żmigrodzki 2022: 155).

All of this influences the richness of the material in particular thematic categories and the completion of them with new units (Batko-Tokarz 2018: 87-88), because the WSJP PAN (as the dictionary of general language) collects lexical units with various features (e.g., scientific, colloquial, expressive, and official features) that will all be included in a given category if they are connected by the same topic. In both dictionaries, not only single words but also discontinuous units (phrasemes and proverbs) are subject to classification. Due to this, the user obtains simultaneous access to both of those groups in the fields. Thematic fields that are created in this way may be useful for specialised users of a language, such as translators, journalists, or authors of various texts who want to extend their vocabulary on a given topic. Additionally, such fields may be useful for both native and non-native language users. Each of them may adjust the access to the dictionary to their needs, and each user may select the linguistic material in a variety of ways. This is enabled by the advanced search options.

Such a dictionary and its material selection options may be very useful for the researchers of a language (Sieradzka-Baziur 2020: 90-96). With the use of

electronic devices, it is easy to carry out varied characteristics of particular categories and their contents and structure as well as quantitative analyses. For example, the field *Życie społeczne* [Social life] dominates in the most developed group in SPJS (*Człowiek* [Human]), which counts as many as 166 categories. The next two places are occupied by those aspects that are related to psychological and spiritual life (Sieradzka-Baziur 2020: 85-86). In the WSJP PAN, the two groups with the largest number of thematically classified units were *Człowiek jako istota psychiczna* [Man as a psychological being] (23,6%) and *Człowiek w społeczeństwie* [Man in society] (22,18% of the units) (Żmigrodzki 2022: 160). Based on such quantitative analyses, interesting conclusions can be drawn on the thematic structure of the vocabulary and comparative analyses can be carried out. It is possible to see which thematic fields or subfields are the most or least numerous (Sieradzka-Baziur 2020: 86-87; Żmigrodzki 2022: 160-164).

4.3 Advanced search

Electronic dictionaries offer various kinds of searching in which modern researchers are increasingly more interested, including various forms of thematic searching (Pastor and Alcina 2010; Camacho-Niño 2021).

A bookmark *Wyszukiwanie zaawansowane* [Advanced search] in the WSJP PAN merits special attention for its exceptional appeal (Figure 6). This is because various search filters may be combined with searching by thematic categories, which gives users an almost unlimited opportunity to arrange thematic fields and search for the material that interests them. The option of searching by thematic criteria in an advanced search looks similar to that on the main page of this dictionary (Option 3: *Wyszukaj według przynależności do kategorii tematycznej* [Search by Thematic Category] (Żmigrodzki 2022: 158-159).

Searching by a specific thematic category can be combined with other options, such as with the option to search by the dictionary's rich set of chronological, expressive, normative, and specialist qualifiers, as well as qualifiers that characterise the vocabulary in terms of the geographical, stylistic, and environmental ranges. Such a searching method can further be combined with searching by the language of origin of a given unit, which enables the semantic analysis of borrowings. Another method of searching is possible by grammar characteristics (i.e., by the types of entries, such as phraseological expressions) or by specified parts of speech. In this way, users may gain access to only the colloquial and expressive vocabulary that is typical of a given specialisation, for example, or to nouns or phraseological expressions that belong to a given thematic category (Batko-Tokarz 2018, 89-91; Żmigrodzki 2022: 158-159). This is a valuable supplementation of the above-described thematic dictionaries of various kinds.

The WSJP PAN has recently added another search option based on the date of the appearance of a given form in Polish (based on the chronologicalising of dictionary entries in the WSJP PAN). This makes it possible to see, for example, how the vocabulary of a specific stage (such as the beginning of the 21st century) is classified thematically and, thus, in which fields the greatest changes

have taken place in this respect. The advanced search, which allows one to search thematic fields and classify the vocabulary that is collected in them according to various criteria, seems to be an invaluable tool that allows one to tailor the thematic dictionary to the diverse needs of its users and, in fact, transform it into the various types of thematic dictionaries that were described above.

Ustawienia
zaawansowanego wyszukiwania

Wpisz szukane słowo... **Szukaj** ↻

1a Wyszukaj hasła (znaczenia) zawierające podany ciąg znaków:
w formie hasłowej w definicji w cytacie w połączeniach

1b Przy porównaniu:
nie uwzględniaj znaków przestankowych i nawiasów
uwzględniaj znaki przestankowe i nawiasy

2 Wyszukiwanie według kwalifikatorów (kwalifikatory):

3 Wyszukaj według przynależności do kategorii tematycznej (Kategoria tematyczna)
[dowolna kategoria] [dowolna podkategoria] [dowolna pozycja]

4 Wyszukiwanie według języka pochodzenia (Pochodzenie):

5 Wyszukiwanie według charakterystyki gramatycznej

6 Wyszukiwanie według chronologizacji
Stulecie: nie określono określono
Rok: Od: Do:

Figure 6: Advanced search in the WSJP PAN

5. Summary

Researchers often indicate the centuries-long domination of the alphabetical arrangement in lexicography (McArthur 1998: 150; Kawalya and De Schryver 2013: 174-177; Kay 2015: 54). Although the same is true from Polish lexicography, the results that were collected in this article show that, in previous centuries as well as in recent decades, many thematic dictionaries have emerged in Polish. Polish scientific research is increasingly focusing on onomasiological and thematic lexicography, which is related to the growing interest in these types of dictionaries in world linguistics as well (Kawalya and De Schryver 2013: 178-180). This phenomenon is reflected in Polish electronic dictionaries that enable the new and definitely more effective functioning of this method and overcoming its limitations. Such dictionaries seem to be the lexicons of the future — especially because they are very useful and their functioning in the electronic form is very promising. In Slavic lexicography, a revival of the research on thematic dictionaries can also be observed in the electronic format; this is clearly related to new lexicographic projects in this area. This is evident from a thesaurus of the Czech language that was published at the beginning of the 21st century, which maintained the format of Roget's thesaurus (Klégr 2007). In the following years, a Czech online onomasiological dictionary and database (<https://beta.najdislovo.cz/>) was developed (Obstová et al. 2022) on the basis of this thesaurus and another Czech thematic dictionary (Haller 1969–1986). Also in Croatian lexicography, the Croatian thesaurus project was described (cf. Nikolić-Hoyt 2004), and now it is a project of the somatic thesaurus for this language (Nikolić-Hoyt 2020). Furthermore, one can find descriptions of projects focused on the compilation of thematic dictionaries for various types of vocabulary (cf. Bihday 2018; Nenasheva 2024).

Endnotes

1. See Batko-Tokarz (2015; 2019: 21-32) for a detailed description of the perception of these terms in European and Polish lexicography.
2. In a dictionary of Slavic linguistic terminology (Jedlička 1977: 222-223), the term *thematic* dominates, but other terms (sometimes several of them) also appear for a given language. The word *ideographical* is popular — especially in the Russian tradition (Burkhanov 1999). In older works, the following names were used: *subject dictionary* — *słownik rzeczowy* (Polish), and *věcný slovník* (Czech). In newer ones terms such as *conceptual: pojęciowy* (Pol.) and *pojmový* (Cz.) are also found.
3. In a narrow understanding, a thematic dictionary is identified only with pedagogical dictionaries that are to support language learning (Burkhanov 2010: 242-243).
4. All of those dictionaries have been characterised in detail against the broader background of the Polish linguistic thought (Batko-Tokarz 2019: 39-127).
5. This was prepared in the late 1960s, but communist censorship prevented its publication because the topics reflected a student perspective that differed from the official state propaganda. The dictionary was only published after the fall of communism in Poland.

6. The one part of the dictionary, arranged in alphabetical order, has been provided with subject indexes modeled on McArthur's (1981) dictionary.
7. In the middle of 2022, however, one of the publishing houses made it possible for users to download the electronic version of the *Słownik tematyczny polsko-ukraiński* [Polish-Ukrainian Thematic Dictionary] (Kononenko, Mytnik and Wasiak 2015). The dictionary was dedicated to Ukrainian refugees who were forced to leave their country and move to Poland because of the war.
8. So far, they do not include image or sound elements. This is one of the challenges that Polish electronic lexicography face, and Polish researchers are thinking about ways of implementing them. For example, Biesaga (2017) analysed which thematic fields were illustrated most often and least often in selected dictionaries by using the thematic scheme applied in the WSJP PAN. Furthermore, she pointed out that it was justified to introduce images that were connected with the thematic division of the vocabulary.
9. This was described in the example of the entry *zdechnąć* [die]. In the WSJP PAN, seven meanings are separated (<https://wsjp.pl/haslo/podglad/35274/zdechnac>) with references to (1) the deaths of animals, and (2) the deaths of human beings, (3) to the functioning of man in the physical aspect (tiredness), and (4) in the psychological aspect (boredom), but also, respectively, with reference (5) to plants, when they wither, (6) to machines and devices when they stop working, and (7) to non-material phenomena that which disappear. Thematic classification emphasises this polysemy as each meaning has been classified in a different way. This is helpful especially to less proficient users of Polish, especially because meanings 3 through 7 are colloquial. In the SPJS, two meanings are distinguished in the entry *zdechnąć*, which receive different thematic classifications (death of people, and death of animals) (<https://spjs.ijppan.pl/haslo/index/21523/9978>).
10. Work on the WSJP PAN is still ongoing, and these numbers should be higher now.

References

Printed dictionaries

- Anusiewicz, J. and J. Skawiński. 1996. *Słownik polszczyzny potocznej*. Warszawa: Wydawnictwo PWN.
- Artomiusz, P. 1591. *Nomenclator selectissimas rerum appellationes tribus linguis, Latina, Germanica, Polonica explicatas indicans*. Kraków.
- Bartmiński, J. and S. Niebrzegowska-Bartmińska (Eds.). 1996–2024. *Słownik stereotypów i symboli ludowych*. Lublin: Wydawnictwo Uniwersytetu Marii Curie-Skłodowskiej.
- Haller, J. (Ed.). 1969–1986. *Český slovník věcný a synonymický 1–3* and *Rejstřík*. Praha: Státní Pedagogické Nakladatelství.
- Hallig, R. and W. von Wartburg. 1952. *Begriffssystem als Grundlage für die Lexikographie*. Berlin: Akademie Verlag.
- Kaczmarek, L., T. Skubalanka and S. Grabias. 1994. *Słownik gwary studenckiej*. Lublin: Wydawnictwo UMCS.
- Kita, M. and E. Polański. 2002. *Słownik tematyczny języka polskiego*. Łódź: Wydawnictwo Literatura.
- Klégr, A. 2007. *Tezaurus jazyka českého. Slovník českých slov a frází souznačných, blízkých a příbuzných*. Praha: Nakladatelství lidové noviny.

- Kononenko, I., I. Mytnik and E. Wasiak.** 2015. *Słownik tematyczny polsko-ukraiński*. Warszawa: Wydawnictwo PWN.
- Kucała, M.** 1957. *Porównawczy słownik trzech wsi małopolskich*. Wrocław: Zakład Narodowy im. Ossolińskich.
- Kurzowa, Z.** 2020. *Ilustrowany słownik podstawowy języka polskiego wraz z indeksem pojęciowym wyrazów i ich znaczeń*. Kraków: Universitas.
- Lebda, R.** 2009. *Wielki słownik frazeologiczny*. Kraków: Krakowskie Wydawnictwo Naukowe.
- McArthur, T.** 1981. *Longman Lexicon of Contemporary English*. Essex: Longman.
- Murmeliusz, J.** 1526. *Dictionarius Ioannis Murmellii variarum rerum*. Kraków.
- Mymer, F.** 1528. *Dictionarium trium linguarum: Latinae: Teutonicae et Polonicae ...* Kraków.
- Pelcowa, H.** 2012–2024. *Słownik gwar Lubelszczyzny*. Lublin: Wydawnictwo UMCS.
- Seretny, A.** 2008. *A co to takiego? Obrazkowy słownik języka polskiego*. Kraków: Universitas.
- Słowarz** (so-called). Anonymous author. 1532. *Dictionarius seu nomenclatura quatuor linguarum*. Kraków.
- Zawiliński, R.** 1926. *Dobór wyrazów. Słownik wyrazów bliskoznacznych i jednoznacznych. Do praktycznego użytku ułożył Roman Zawiliński*. Kraków: "Poradnik Językowy". Accessed at: <https://rcin.org.pl/dlibra/show-content/publication/edition/36954?id=36954> (23 April 2005)

Online dictionaries

- SPJS** — Sieradzka-Baziur, B. (Ed.). *Słownik pojęciowy języka staropolskiego*. Accessed at: <https://spjs.ijp.pan.pl/spjs/strona/kartaTytuLOWa> (16 November 2025)
- WSJP PAN** — Żmigrodzki, P. (Ed.). 2006–. *Wielki słownik języka polskiego PAN*. Accessed at: <https://wsjp.pl/> (16 November 2025)

Other literature

- Balalaieva, O.** 2021. Класифікації Електронних Словників [Classifications of Electronic Dictionaries], *Міжнародний Філологічний Часопис* [International Philological Journal] 12(2): 60–63. <https://doi.org/10.31548/philolog2021.02.060>
- Batko-Tokarz, B.** 2008. Tematyczny podział słownictwa w Wielkim słowniku języka polskiego. Żmigrodzki, P. and R. Przybylska (Eds.). 2008. *Nowe studia leksykograficzne 2*: 31–48. Kraków: Lexis. Accessed at: <http://ijp.pan.pl/publikacje-elektroniczne/nowe-studia-leksykograficzne-t-2> (1 April 2025)
- Batko-Tokarz, B.** 2015. Słowniki o układzie hierarchicznym — próba uporządkowania pojęć. *Język Polski* 95(5): 406–417. <https://doi.org/10.31286/JP.95.5.2>
- Batko-Tokarz, B.** 2018. Klasyfikacja tematyczna w WSJP PAN. Żmigrodzki, P., M. Bańko, B. Batko-Tokarz, J. Bobrowski, A. Czelakowska, M. Grochowski, R. Przybylska, J. Waniakowa and K. Wegrzynek (Eds.). 2018. *Wielki słownik języka polskiego PAN. Geneza, koncepcja, zasady opracowania*: 79–92. Kraków: Libron. Accessed at: <https://rcin.org.pl/dlibra/publication/edition/68631?id=68631&from=publication> (10 April 2025)
- Batko-Tokarz, B.** 2019. *Tematyczny podział słownictwa współczesnego języka polskiego — teoria, praktyka, leksykografia*. Kraków: Wydawnictwo Uniwersytetu Jagiellońskiego. Accessed at: <https://ruj.uj.edu.pl/entities/publication/f52bfba-4e31-4246-bd1b-f222891e8d06> (12 April 2025)

- Biesaga, M.** 2017. Dictionary Tradition vs. Pictorial Corpora: Which Vocabulary Thematic Fields Should Be Illustrated? *Lexikos* 27: 132-151.
<https://doi.org/10.5788/27-1-1397>
- Bihday, M.** 2018. Representation of Lexical-Semantic Classification of the Ukrainian Verbs in the Form of Ideographic Dictionary. *Linguistic Studies* 36: 44-49.
<https://doi.org/10.31558/1815-3070.2018.36.6>
- Burkhanov, I.** 1999. *Linguistic Foundations of Ideography: Semantic Analysis and Ideographic Dictionaries*. Rzeszów: Wydawnictwo Wyższej Szkoły Pedagogicznej.
- Burkhanov, I.** 2010. *Lexicography: A Dictionary of Basic Terminology*. Rzeszów: Wydawnictwo Uniwersytetu Rzeszowskiego.
- Camacho-Niño, J.** 2021. Estructuras de acceso y remisión en los diccionarios del siglo XXI. Moreno Moreno, M.A. and M. Torres Martínez (Eds.). 2021. *Estudios del léxico en el ámbito universitario del siglo XXI*: 125-148. Barcelona: Editorial Octaedro S.L. Accessed at:
https://www.researchgate.net/publication/352996840_Estructuras_de_acceso_y_remision_en_los_diccionarios_del_siglo_XXI (1 April 2025)
- Daković, S.** 2011. Słowniki tematyczne w Polsce, Serbii i Chorwacji. *Südslavistik online — zeitschrift für südslavische Sprachen und Literaturen und Kulturen* 3. Dvojezična leksikografija bosanskog/hrvatskog/srpskog jezika: 33-52. Accessed at:
<http://www.suedslavistik-online.de/03/03.pdf> (1 April 2025).
- Ferrett, E. and S. Dollinger.** 2021. Is Digital Always Better? Comparing Two English Print Dictionaries with Their Digital Counterparts. *International Journal of Lexicography* 34(1): 66-91.
<https://doi.org/10.1093/ijl/ecaa016>
- França, P.C.** 2010. *Onomasiological Dictionaries and Ontologies*. Dykstra, A. and T. Schoonheim (Eds.). 2010. *Proceedings of the XIV Euralex International Congress, Leeuwarden, 6-10 July 2010*: 1291-1298. Ljouwert: Fryske Akademy. Accessed at:
http://www.euralex.org/elx_proceedings/Euralex2010/121_Euralex_2010_9_FRANCA_Onomasiological%20dictionaries%20and%20ontologies.pdf (8 April 2025)
- Granger, S. and M. Paquot (Eds.).** 2012. *Electronic Lexicography*. Oxford: Oxford University Press.
- Greń, Z.** 2023. Klasyfikacja tematyczna słownictwa w leksykografii czeskiej jako źródło do badania zmian w językowym obrazie świata. Falski, M. and E. Kaczmarek-Zglejszewska (Eds.). 2023. *Czechy przez naukę języka do rozumienia kultury. Księga dedykowana Teresie Piotrowskiej-Malek*: 15-43. Kraków: Libron.
- Gruszczyński, W. and Z. Saloni.** 2013. From Multilingual to Monolingual Dictionaries. A Historical Overview of Polish Lexicography. *Studies in Polish Linguistics* 8(4): 205-227.
<https://doi.org/10.4467/23005920SPL.13.010.1691>
- Hartmann, R.R.K.** 2005. Onomasiological Dictionaries in 20th-Century Europe. *Lexicographica. International Annual for Lexicography* 21: 6-19.
<https://doi.org/10.1515/9783484604742.6>
- Hartmann, R.R.K. and G. James.** 2002. *Dictionary of Lexicography*. London/New York: Routledge.
- Hüllen, W.** 2006. *English Dictionaries, 800-1700. The Topical Tradition*. Oxford: Oxford University Press.
- Jedlička, A. (Ed.).** 1977. *Slovník slovanské lingvistické terminologie = Словарь славянской лингвистической терминологии = Dictionary of Slavonic Linguistic Terminology*. Praha: Academia nakladatelství Československé akademie věd Praha.

- Kawalya, D. and G.-M. de Schryver.** 2013. Introducing a New Lexicographical Model: AlphaConceptual⁺ (and How it Could Be Applied to Dictionaries for Luganda). *Lexikos* 23: 172-200.
<https://doi.org/10.5788/23-1-1210>
- Kononenko, I. and I. Mytnik.** 2017. Potencjal teoretyczny i praktyczny dwujęzycznego słownika tematycznego. Borys, O., M. Jeż, A. Samadowna and M. Saniewska (Eds.). 2017. *Szkice językowe i literacko-kulturowe*: 40-57. Warszawa: Katedra Ukrainistyki, Uniwersytet Warszawski. Accessed at: <https://open.icm.edu.pl/handle/123456789/19237> (10 April 2025)
- Koutny, I.** 2005. A Corpora Assisted Multilingual Thematic Dictionary. *Archives of Control Sciences* 15(4): 573-582.
- Koutny, I.** 2012. From Semantic Networks to Dictionary Structure. *Język. Komunikacja. Informacja. Language. Communication. Information* 7: 115-128. Accessed at: <https://repozytorium.amu.edu.pl/server/api/core/bitstreams/61eb6b33-8db3-4f17-8023-25e811b2968d/content>
- Kozioł-Chrzanowska, E.** 2017. What Do Users of General Electronic Monolingual Dictionaries Search For? The Most Popular Entries in the *Polish Academy of Sciences Great Dictionary of Polish*. Kosem, I., C. Tiberius, M. Jakubiček, J. Kallas, S. Krek and V. Baisa (Eds.). 2017. *Electronic Lexicography in the 21st Century. Proceedings of eLex 2017 Conference, Leiden, the Netherlands, 19–21 September 2017*: 202-220. Brno: Publisher: Lexical Computing CZ s.r.o., Brno, Czech Republic. Accessed at: <https://elex.link/elex2017/wp-content/uploads/2017/09/paper12.pdf>
- McArthur, T.** 1998. *Living Words: Language, Lexicography and the Knowledge Revolution*. Exeter: University of Exeter Press.
- Nenasheva, L.V.** 2024. «Тематический словарь архангельских говоров» и электронная система словаря ["Thematic Dictionary of Arkhangelsk Dialects" and the Electronic Dictionary System]. *Philology. Theory & Practice* 17(9): 3376-3383.
<https://doi.org/10.30853/phil20240478>
- Nikolić-Hoyt, A.** 2004. *Konceptualna leksikografija: prema tezaurusu hrvatskoga jezika*. Zagreb: Hrvatska sveučilišna naklada.
- Nikolić-Hoyt, A.** 2020. Iz riznice Benešićeva rječnika i Somatskoga tezaurusa hrvatskoga jezika. *Filologija* 75: 47-65.
<https://dx.doi.org/10.21857/y54jofkgqm>
- Obstová, Z., O. Tichý and A. Klégr.** 2022. Projekt Lexikálně-sémantické databáze češtiny (LSD-Czech): uživatelský pojmový slovník a online databáze. *Studie z Aplikované Lingvistiky* 13: 86-98. Accessed at: https://studiezaplikovanelingvistiky.ff.cuni.cz/wp-content/uploads/sites/19/2023/01/Zora_Obstova_-_Ondrej_Tichy_-_Ales_Klegr_86-98.pdf (28 April 2025)
- Pastor, V. and A. Alcina.** 2010. Search Techniques in Electronic Dictionaries: A Classification for Translators. *International Journal of Lexicography* 23(3): 307-354.
<https://doi.org/10.1093/ijl/ecq015>
- Rembiszewska, D.K. (Ed.).** 2016. *Słowiańskie słowniki gwarowe*. Warszawa-Łomża: Łomżyńskie Towarzystwo Naukowe im. Wagów.
- Sękowska, E.** 2020. A Thematic Dictionary as a Source of Cultural Competence (Using the Example of a Chinese-Polish Dictionary). *Acta Universitatis Lodzianis. Kształcenie Polonistyczne Cudzoziemców* 27: 339-348.
<https://doi.org/10.18778/0860-6587.27.19>

- Sieradzka-Baziur, B.** 2020. *Onomazjologiczny słownik online. Metodologia, zawartość, wykorzystanie w badaniach na przykładzie "Słownika pojęciowego języka staropolskiego"*. Kraków: Instytut Języka Polskiego Polskiej Akademii Nauk. Accessed at: <https://rcin.org.pl/dlibra/publication/270658/edition/233563> (12 April 2025)
- Sierra, G.** 2000. *The Onomasiological Dictionary: A Gap in Lexicography*. Heid, U., S. Evert, E. Lehmann and C. Rohrer (Eds.). 2000. *Proceedings of the Ninth EURALEX International Congress, EURALEX 2000, Stuttgart, Germany, August 8th–12th, 2000*: 223-235. Stuttgart: Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart. Accessed at: http://www.euralex.org/elx_proceedings/Euralex2000/027_Gerardo%20SIERRA_The%20onomasiological%20dictionary_a%20gap%20in%20lexicography.pdf (12 April 2025)
- Żmigrodzki, P.** 2008. *Słowo — słownik — rzeczywistość. Z problemów leksykografii i metaleksykografii*. Kraków: Lexis.
- Żmigrodzki, P.** 2018. Methodological Issues of the Compilation of the *Polish Academy of Sciences Great Dictionary of Polish*. Čibej, J., V. Gorjanc, I. Kosem and S. Krek (Eds.). 2018. *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts, 17–21 July 2018, Ljubljana, Slovenia*: 209-219. Ljubljana: Ljubljana University Press, Faculty of Arts. Accessed at: <https://euralex.org/publications/methodological-issues-of-the-compilation-of-the-polish-academy-of-sciences-great-dictionary-of-polish/>
- Żmigrodzki, P.** 2022. *Wielki słownik języka polskiego PAN jako słownik tematyczny. Studia Językoznawcze. Synchroniczne i diachroniczne aspekty badań polszczyzny* 21: 153-166. Accessed at: <https://wnus.edu.pl/sj/pl/issue/1252/article/19943/>

Spelling Variations in *A Dictionary of Setswana Personal Names*

Thapelo J. Otlogetswe (otlogets@ub.ac.bw)

(<https://orcid.org/0000-0003-3887-2982>)

and

Goabilwe Ramaeba (ramaebagn@ub.ac.bw)

(<https://orcid.org/0000-0002-2379-9703>)

Faculty of Humanities, University of Botswana, Gaborone, Botswana

Abstract: This paper investigates spelling variation in *A Dictionary of Setswana Personal Names* (Otlogetswe and Ramaeba 2024), addressing a gap in existing research which has previously focused on general-purpose monolingual Setswana dictionaries (Otlogetswe 2015) but not specialized dictionaries such as the one on personal names. The study draws on orthographic guidelines outlined in the *Standard Orthography of 1981* (Ministry of Education 1981) and *Mokwalo o o Lolameng wa Setswana* (Chebanne et al. 2008), while also tracing the historical development of Setswana orthography to early works by Lichtenstein (1928–1930; 1973) and Livingstone (1875). Using textual analysis, the paper critically examines dictionary headwords to identify and categorize spelling variants. The findings reveal numerous inconsistencies in the spelling of Setswana personal names, influenced by factors such as the coexistence of older and newer orthographic standards, dialectal variation across Setswana-speaking regions, and underlying morphological and phonological processes.

Keywords: SPELLING VARIATION, A DICTIONARY OF SETSWANA PERSONAL NAMES, SETSWANA, ORTHOGRAPHY

Opsomming: Spellingvariasie in *A Dictionary of Setswana Personal Names*.

In hierdie artikel word die spellingvariasie in *A Dictionary of Setswana Personal Names* (Otlogetswe en Ramaeba 2024) ondersoek, om sodoende 'n gaping te vul in die bestaande navorsing wat voorheen gefokus het op algemene Setswana-woordeboeke (Otlogetswe 2015), maar nie op gespesialiseerde woordeboeke soos die een oor persoonsname nie. Die studie steun op ortografiese riglyne, soos uiteengesit in die *Standard Orthography of 1981* (Ministerie van Onderwys 1981) en *Mokwalo o o Lolameng wa Setswana* (Chebanne et al. 2008), terwyl die historiese ontwikkeling van Setswana-ortografie na vroeë werke deur Lichtenstein (1928–1930; 1973) en Livingstone (1875) teruggespoor word. Met behulp van tekstuele analise ondersoek die artikel woordeboektrefwoorde krities om spellingvariante te identifiseer en te kategoriseer. Die bevindings toon talle inkonsekwentheid in die spelling van Setswana-persoonsname wat beïnvloed word deur faktore soos die gelyktydige voorkoms van ouer en nuwer ortografiese standaarde, dialektiese variasie in Setswana-sprekende streke, en onderliggende morfologiese en fonologiese prosesse.

Sleutelwoorde: SPELLINGVARIASIE, A DICTIONARY OF SETSWANA PERSONAL NAMES, SETSWANA, ORTOGRAFIE

1. Introduction and background

Spelling variations are ubiquitous in natural languages, and they have become an essential element of any dictionary since they capture regional nuances, historical contexts, and the dynamic nature of language. In a globalised world, people encounter texts from various regions characterised by multiple variations. Capturing spelling variations can, therefore, go a long way to help readers understand and recognise different spellings, as such reducing confusion in text reception. For instance, English is spoken and used in many countries with different spelling conventions. Knowing that "color" and "colour" are the same word but used in different regions can prevent misunderstandings and minimize confusion. For language learners, encountering different spellings can be challenging. Dictionaries that include variations provide a comprehensive learning tool, helping learners become familiar with regional differences and enhancing their overall language proficiency.

Similarly, Africa is home to a multitude of languages and dialects, each with its own unique spelling conventions. The Centre for Advanced Studies of African Society (CASAS) leads the work to harmonize and standardize language orthographies in Africa to minimise the existence of competing orthographies of the same language (Chebanne et al. 2008, Prah 1998). A typical case study on how to handle spelling variations in an African language dictionary has been discussed by Otlogetswe (2015) looking at variation in the Setswana monolingual dictionary *Tlhalosi ya Medi ya Setswana* (Otlogetswe 2012). Setswana has borrowed many words from English and Afrikaans, and these words are often spelled in different ways because they are borrowed as they are heard. For example, the word "computer" has been borrowed as *khompiutara*, *khomphiutara*, *khomputara*, *komputara*, *khomputa*, or *khompiuta*. Since there is no language board that Setswana spellings of borrowed words in Botswana, variants remain in the language. Otlogetswe argues that spelling variants should be included in general monolingual dictionaries and that the way they are handled should be informed by frequency information from corpus data.

Dictionaries like the *African Pidgin Dictionary* (see www.pegeen.com) list spelling variations to ensure users from different regions can find the relevant spelling. For instance, a word might have different spellings in Nigerian, Ghanaian, and Cameroonian pidgin. By including these variations, dictionaries capture and document the rich linguistic diversity of a language. This is particularly important for preserving regional dialects and cultural identities. To help users understand the appropriate context for each spelling variant, African dictionaries can use usage labels and contextual notes (Otlogetswe 2014). These labels might indicate whether a spelling is specific to a particular country or region. For example, the South African National Lexicography Units (SANLU) provide detailed notes on the usage of words in different South African languages, helping users navigate the regional differences. This approach ensures that users can select the correct spelling based on their specific needs and contexts.

Additionally, language evolves over time, and spelling variations often reflect

historical changes (Murray 1996). Including archaic or historical spellings in dictionaries helps users understand and trace the development of the language and provides context for older texts. For example, knowing that "jail" was once spelled "gaol" can be useful for reading and understanding historical documents. Some African dictionaries include historical or archaic spellings to provide a comprehensive view of a word's evolution. This is particularly important for languages with rich oral traditions and historical texts. By including these spellings, dictionaries help users understand the historical context and development of the language. For instance, older spellings of Setswana words might be included to aid in the study and appreciation of historical documents.

Spelling variations also support accurate and contextual usage (Motzafi-Haller et al. 2006). Some words have different spellings depending on their context. For example, "dialog" is often used in computing contexts, while "dialogue" is preferred for conversations between people. Dictionaries that include these variations help users choose the correct spelling for their specific context. Lexicographers make editorial decisions about which variants to include in a dictionary based on the word's usage frequency and relevance. For example, if a particular spelling is commonly used in contemporary literature or media, it might be prioritised in the dictionary. This ensures that the dictionary remains relevant and useful for modern users. For instance, the inclusion of both "doughnut" and "donut" in dictionaries reflects their usage in different contexts. The "donut" spelling has been widely used in the USA from the mid-20th century, while the "doughnut" spelling has been used consistently elsewhere.

The inclusion of spelling variations in a dictionary can also go a long way in preserving the cultural heritage of a linguistic group (Scontras et al. 2015). Including spelling variations for indigenous and local terms helps preserve cultural heritage and ensures that these terms are recognised and respected. This is especially important in African dictionaries, where capturing the diversity of languages and dialects is crucial. Proper nouns and brand names often have unique spellings. Including these in dictionaries helps users understand that while these spellings are correct for specific entities, they are not generalisable to other contexts. This distinction is important for accurate communication and brand recognition.

The inclusion of spelling variation also has practical applications. In the digital age spellcheckers and autocorrect tools rely on comprehensive dictionaries (Mulingi 2023). Including spelling variations ensures that these tools can recognise and correct different spellings, enhancing their effectiveness and user experience. As English continues to be a global lingua franca, understanding and recognising spelling variations is essential for effective communication across different regions. Dictionaries that include these variations support global communication by providing a common reference point (<https://blog.spellquiz.com/spelling-variants/>).

The treatment of spelling variants in Setswana dictionaries provides a useful case study. The dictionaries *Thanodi ya Setswana* (Kgasa and Tsonope 1995), *Thanodi ya Setswana* (Mareme 2007), and *Tlhalosi ya Medi ya Setswana* (Otlogetswe 2012) have all dealt with spelling variation in different ways. These dictionaries reflect

the influence of regional dialects and the evolution of the language over time.

As previously indicated, the *African Pidgin Dictionary* is another example of how spelling variations are handled. This dictionary includes words from various Pidgin English dialects spoken across Africa, such as Nigerian Pidgin and Ghanaian Pidgin (<https://peegeen.com/>). It lists multiple spellings for words and provides contextual notes to help users understand their regional differences. This approach ensures that the dictionary is a comprehensive resource for speakers of different Pidgin English dialects. The Shona monolingual dictionary developed by the African Languages Research Institute (ALRI) at the University of Zimbabwe grapples with dialectal diversity in Shona (Chabata 2003). It addresses tone marking, sense ordering, and synonym treatment to reflect dialectal nuances. However, representativeness remains a challenge, as some dialects are underrepresented.

In practical lexicography, cross-references are commonly used to guide users from one spelling variant to another (Prinsloo and Van Graan 2021). If a user looks up a word in one spelling, the dictionary might direct them to another variant with a note explaining the regional or contextual differences. This method ensures that users find the information they need, regardless of the spelling they initially searched for. For example, a dictionary might list "catalogue" and direct users to "catalog" with a note on regional preferences.

African dictionaries, however, face unique challenges due to the continent's linguistic diversity and the dynamic nature of its languages. One challenge is the standardisation of spelling in languages that have multiple dialects and regional variations (Baugh 2009). To address this, lexicographers often collaborate with local language experts and communities to ensure that the dictionary accurately reflects current usage and captures variations as synonyms in dictionaries (Chabata 2003). Innovations in technology have also played a significant role in developing African dictionaries. Digital platforms allow for more frequent updates and the inclusion of multimedia elements such as audio pronunciations and video clips. These features enhance the usability of the dictionary and provide a richer linguistic resource for users. African dictionaries handle spelling variations by including regional variants, using usage labels and contextual notes, providing cross-references, including historical spellings, making editorial decisions based on usage, and addressing proper nouns and indigenous terms. These strategies help users navigate the complexities of African languages, ensuring they can find and use the correct spelling for any given context. By reflecting the linguistic diversity and cultural richness of the continent, African dictionaries play a vital role in education and language preservation.

2. Methodology

This paper analyses spelling variations found in names that are headwords in *A Dictionary of Setswana Personal Names*. The names are analysed based on *Standard Orthography of 1981* (Ministry of Education 1981) and *Mokwalo o o lolameng wa*

Setswana (Chebanne et al. 2008). While the historical roots of Setswana orthography may be traced to Lichtenstein (1928) and Livingstone (1875), the Ministry of Education (1981) and Chebanne et al. (2008) present recent rules to Setswana spelling in Botswana. The study uses a textual analysis method to critically assess the dictionary headwords for spelling variants. The results show that there are numerous inconsistencies in the spelling of Setswana personal names. The spellings are motivated by how speakers pronounce the names and not necessarily how they ought to be spelt, according to the standard Setswana orthography. The variations may be clustered into four categories of: (a) vowel variations, (b) consonantal variations, and (c) dialect influenced variations. The final section of the paper deals with names that use discontinued consonants.

Results and discussion

Spelling variations in African personal names are a widespread and multifaceted issue, shaped by linguistic diversity, colonial legacies, orthographic inconsistencies, and sociocultural dynamics. They are a consequence of variation at consonantal and vowel levels as demonstrated in the following sections.

3. Vowel variations in Setswana names

Setswana has seven vowels. These are i, e, ε, u, ɔ, u, a (DALL 1999: 17). The seven vowels may be represented as in Table 1:

Height	Localisation	
	Front	Back
Close	i	U
Half-close	e	ɔ
Half-Open	ε	ɔ
Open	A	

Table 1: Setswana vowels

Orthographically, Setswana does not mark a distinction between /ε/ and /e/. They are both represented by [e] in normal written texts, except in the Botswana lower primary school syllabus where /ε/ is represented by [e] with a circumflex as [ê]. Equally, Setswana orthography does not mark a distinction between /u/ and /ɔ/ except in the same syllabus where the /ɔ/ is represented by [o] with a circumflex as [ô]. From the Botswana upper primary syllabus and beyond, written Setswana does not mark a distinction between /u/ and /ɔ/ and between

/ɛ/ and /e/ (Sebina and Otlogetswe 2023a and b). Since Setswana does not mark the /ɛ, e/ and /o, ɔ/ distinction orthographically, there are no names that are spelt with [ê] or [ô]. Just as in common writing, Setswana names use the five orthographic vowels, a, e, i, o, u. This discussion of vowels is relevant to the vowel variations that follow below since several names do not follow standard orthography where vowels are concerned.

3.1 Names that use [i] instead of [e]

Name spelling variation in Setswana is found in names which in their articulation speakers use the half-close front vowel /e/, instead of the close front vowel /i/. This practice is then reflected in the spelling of the same name with one spelt with an /e/ while the same name is also spelt with /i/. Some speakers therefore spell their names as *Baaitsi* instead of *Baaitse*, *Matshidiso* instead of *Matshediso*, *Balibi* instead of *Balebi*, *Kelitile* instead of *Keletile*, *Baipidi* instead of *Baipedi*, *Baoliki* instead of *Baoleki*, *Bafiti* instead of *Bafeti* and *Goitsimodimo* instead of *Goitsemodimo*. This may be explained in terms of close vowel harmony where the preceding vowel harmonises in height with the following one. The /e/ is therefore pronounced as an /i/. The correct spelling of the name is traceable to the verb from which the name is derived. *Baaitse* is from *itse* "know", *Matshediso* from *tshela* "live", *Balebi* from *leba* "look", *Keletile* from *leta* "wait", *Baipedi* from *ipela* "celebrate", *Baoleki* from *leka* "try/test", *Bafeti* from *feta* "pass" and *Goitsemodimo* from *itse* "know".

The extracts from the dictionary in Table 2 illustrate how the dictionary treats both the wrong and the correct spellings.

Wrong spelling entry	Correct spelling entry
Balibi (the correct spelling is <i>Balebi</i> from the verb <i>leba</i> "look") <i>M</i> or <i>F</i> those who look.	Balebi <i>M</i> onlookers; those who are looking.
Baliki (the correct spelling is <i>Baleki</i> from the verb <i>leka</i> "try; attempt") <i>M</i> those who try; those who test someone or something.	Baleki <i>F</i> the trying ones.
Balisi (the correct spelling is <i>Balesi</i> from the verb <i>lesa</i> "leave; let go") <i>M</i> those who leave something.	Balesi <i>M</i> those who leave something.
Goitsimodimo (the correct spelling is <i>Goitsemodimo</i>) <i>M</i> it is God who knows.	Goitsemodimo <i>M</i> it is God who knows.

Table 2: Names that incorrectly use [i] instead of [e]

The wrong spelling entry is cross-referenced to the correct spelling entry though the correct spelling entry is not cross-referenced to the wrong spelling. The cor-

rect spelling entry could be improved by cross-referencing the wrong spelling entry to make the user aware of the misspelt version in the following manner:

Balebi *M* on lookers; those who are looking
(usually misspelt as *Balibi*).

3.2 Names that use [e] instead of [i]

Other spelling variations are introduced in a reversal of the situation discussed in 3.1, where name givers use the half-close front vowel and write /e/ where the close front vowel /i/ should be used, thereby violating the agentive rule that derives -i verbs. Cole (1955: 113) has demonstrated that one way of forming Setswana nouns is from verb stems, by prefixing the class prefixes, while the final vowel of the stem is changed to the agentive marker -i (e.g. mo- + dira + -i = modiri). They, therefore, spell their names as *Gabaiphewe* instead of *Gabaiphiwe*, *Molapise* instead of *Molapisi*, *Baitshenyetse* instead of *Baitshenyetsi*, *Molebatse* instead of *Molebatsi*, *Kebaitsele* instead of *Kebaitsile*, and *Modiege* instead of *Modiegi*. This is possibly because the name givers gave the name by ear as in hypothesised in 3.2 and in their hearing, the /i/ sounded like an /e/. Table 3 further illustrates this:

Wrong spelling entry	Correct spelling entry
Molapise (the correct spelling is <i>Molapisi</i>) <i>M</i> one that tires others.	Molapisi (see <i>Molapise</i>) <i>M</i> one that tires others.
Molebatse (the correct spelling is <i>Molebatsi</i>) <i>M</i> one that makes you forget.	Molebatsi <i>M</i> one that makes you forget.
Modiege (the correct spelling is <i>Modiegi</i>) <i>F</i> one that delays.	Modiegi <i>F</i> one that delays.
Baitshenyetse (the correct spelling is <i>Baitshenyetsi</i>) <i>M</i> those that spoil things for themselves.	Baitshenyetsi <i>F</i> those that spoil things for themselves.

Table 3: Names that incorrectly use [e] instead of [i]

The wrong spelling entry is cross-referenced to the correct spelling entry to direct the users to the correct-spelt entry. The correct spelling entry is not cross-referenced to the wrong spelling entry. As in 3.1 above, the correct spelling entry could be improved by cross-referencing the wrong spelling entry to make the user aware of the misspelt version in the dictionary like this:

Molapisi (see *Molapise*) *M* one that tires others
(usually misspelt as *Molapise*).

3.3 Names that use double [o] instead of [eo]

There are names that use [oo] where the standard orthography [eo] should be used. Such names are prevalent in the central and southern regions of Botswana. This spelling variation is typical of the pronunciation of central and southern Setswana dialects amongst the Ngwaketse, Rolong and Hurutshe. This spelling change mirrors vowel assimilation in Setswana speech where one vowel influences another — typically across syllable boundaries — so that they become more similar in quality. Vowel assimilation maintains phonotactic harmony and ease of articulation in speech. While this is true in speech, spelling rules should not be violated as they have been in some of the names. Vowel assimilation has therefore influenced the spelling of certain names such that we have *Kooagile* instead of *Keoagile*, *Koobake* instead of *Keobake*, *Koobokile*, instead of *Keobokile*, *Koolebile* instead of *Keolebile*, *Koorapetse* instead of *Keorapetse* and *Koobame* instead of *Keobame*. The sound change from [eo] to [oo] is a consequence of total regressive assimilation where a preceding sound becomes the same as the following one. For example,

Ke + obame = Koobame

This process helps maintain phonotactic harmony and ease of articulation in speech. In this instance, the [e] of [Ke] becomes like [o] of [obame] to result in *Koobame*. Those who spell the names with [oo] instead of [eo], therefore, write as they speak, and do not adhere to the Setswana standard orthography. *A Dictionary of Setswana Personal Names* has captured both these spelling variations as demonstrated in Table 4:

Wrong spelling entry	Correct spelling entry
Kooagile (the correct spelling is <i>Keoagile</i>) <i>M</i> I have built it.	Keoagile <i>M</i> I have built it.
Koobake (the correct spelling is <i>Keobake</i>) <i>M</i> that I should praise Him (God).	Keobake 1. <i>M</i> I should praise Him (God) 2. should I praise Him (God)? 3. (short for <i>Gakeobake</i>) I am not praising Him (God).
Koobame (the correct spelling is <i>Keobame</i>) <i>M</i> I have worshipped; I am bent.	Keobame 1. <i>F</i> I am bending forward 2. I worshipped.
Koobokile (the correct spelling is <i>Keobokile</i>) <i>M</i> or <i>F</i> I have praised Him (God).	Keobokile <i>M</i> I praised Him (God).

Table 4: Names that incorrectly use [oo] instead of [eo]

The wrong spelling entry is cross-referenced to the correct spelling entry though the correct spelling is not cross-referenced to the wrong spelling entry. The correct

spelling entry could be improved by cross-referencing the wrong spelling entry to make the user aware of the misspelt version in the following manner:

Keobame 1. *F* I am bending forward **2.** I worshipped
 (usually misspelt as *Koobame*).

3.4 Names that delete the vowel [e] in the context of the lateral approximant /l/

There is evidence of names where the vowel [e] gets deleted when it occurs between two lateral approximants /l/ as in the word *molelo* "fire", which is then spelt as *mollo*, after deleting the vowel [e], and retaining the second vowel [o]. This then results in the first /l/ becoming a tone-bearing syllabic consonant. This gives rise to names such as *Mollowakgosi* instead of *Molelowakgosi* and *Mollowakgotla* instead of *Molelowakgotla*. While the [e] is elided in speech, such an elision violates standard Setswana orthography (Chebanne et al. 2008: 64) that states that "... tlogelo ya tumanosi fa gare ga bo[l] ga e letlelelwe ..." (an elision of a vowel in between a sequence of the lateral /l/ is not permitted.) Table 5 further illustrates this:

Wrong spelling entry	Correct spelling entry
Mollowakgosi (the correct spelling is <i>Molelowakgosi</i>) <i>M</i> the chief's fire.	Molelowakgosi <i>M</i> the chief's fire.
Mollowakgotla (the correct spelling is <i>Molelowakgotla</i>) <i>M</i> the fire at the traditional court.	Molelowakgotla <i>M</i> the fire at the traditional court

Table 5: [e] elision between laterals

The wrong spelling entry is cross-referenced to the correct spelling entry though the correct spelling is not cross-referenced to the wrong spelling entry. The correct spelling entry could be improved by cross-referencing the wrong spelling entry to make the user aware of the misspelt version in the following manner:

Molelowakgotla *M* the fire at the traditional court
 (usually misspelt as *Mollowakgotla*).

4. Consonant variations in Setswana personal names

Some name variations are a result of a difference in consonantal use for the same part of a name. Before we delve deeper into consonantal differences, first we present the Setswana consonantal inventory. Setswana has 29 phonemic consonants (Batibo 2013 and DALL 1999: 12) as illustrated in Table 6.

	Labial	Alveolar		Post alveolar	Palatal	Velar	Uvular	Glottal
		Central	Lateral					
NASAL	/m/	/n/			/ɲ/	/ŋ/		
PLOSIVE	Unaspirated	/p/ /b/	/t/ /d/			/k/		
	Aspirated	/p ^h /	/t ^h /			/k ^h /	/q ^h /	
AFFRICATE	Unaspirated		/ts/	/tʃ/	/tʃ/ /dʒ/			
	Aspirated		/ts ^h /	/tʃ ^h /	/tʃ ^h /			
FRICATIVE	/f/	/s/		/ʃ/			/χ/	/h/
Trill		/r/						
Approximant	/w/		/l/		/j/			

Table 6: Setswana consonants

Four of these (p^h, t^h, k^h and q^h) are aspirated voiceless plosives, three are non-aspirated voiceless plosives (p, t and k), one voiced plosive /b/, three aspirated voiceless affricates (tʃ^h, tʃ^h, ts^h), three non-aspirated voiceless affricates (ts, tʃ, and ʃ), one voiced affricate (dʒ), five voiceless fricatives (f, s, ʃ, χ, and h), four nasals (m, n, ɲ, and ŋ), and four sonorants (r, l, j and w).

The voiced plosive /d/ is an allophone of /l/ where /d/ is usually followed by close vowel /u/ and /i/ as in *dika* (surround) and *huduga* (relocate) while /l/ is followed by open vowels such as /a/ and /o/ as in *lala* (spend the night) and *lole* (fought) (see Cole 1955: 28). The exception to this rule is borrowed words such as *lithara* (liter) and *dolara* (dollar).

The Setswana consonantal phonemes are represented orthographically with their equivalent forms as demonstrated in Table 7.

Phonetic	Orthography	/ts/	Ts	/ɲ/	ny
/m/	m	/ts ^h /	Tsh	/j/	y
/p/	p	/s/	S	/ŋ/	ng
/p ^h /	ph	/r/	R	/k/	k
/b/	b	/tʃ/	Tl	/k ^h /	kh
/f/	f	/tʃ ^h /	Tlh	/q ^h /	kg
w/	w	/l/	L	/χ/	g
/n/	n	/ʃ/	Tš	/h/	h
/t/	t	/dʒ/	J	Total	29
/d/	d	/tʃ ^h /	Tšh		
/t ^h /	th	/ʃ/	Š		

Table 7: Phonetic consonants and their orthographic forms

A discussion of Setswana consonantal inventory will prove relevant to sections 4.1–4.3 where consonantal variation in names is discussed.

The use of /h/ instead of /f/

Some name spelling variations in Setswana are because the same name is spelt either with a voiceless, labio-dental fricative [f] in some cases, while in other cases it is spelt with the glottal fricative [h]. Using the sound [h] instead of [f] is a common feature of Setswana speech and there is much confusion in written Setswana about where [h] and [f] should be used (Cole 1955: 26). There is, however, a Setswana morphological rule which illustrates that [h] verbs result with [kh] deverbative nouns and [f] verbs result with [ph] deverbative nouns in strengthening (Cole 1955: 40). For instance, the verb *hupela* generates the noun *khupelo* while the verb *fenya* generates the noun *phenyo*. The [h] and [f] spelling confusion are pervasive in personal names. This results in names such as *Bahedile* instead of *Bafedile*, *Bahemedi* instead of *Bafemedi*, *Gaehole* instead of *Gaefole*, *Hetolang* instead of *Fetolang* and *Hetanang* instead of *Fetanang*. One way of determining the correct spelling for each name is therefore to trace it to its original verb. Table 8 further illustrates how these are captured in the dictionary:

Wrong spelling entry	Correct spelling entry
Bahedile (the correct spelling is <i>Bafedile</i> from the verb <i>fela</i> "end; finish") <i>M</i> they are finished.	Bafedile <i>M</i> they are finished.
Bahemedi (the correct spelling is <i>Bafemedi</i> from the verb <i>fema</i> "defend") <i>M</i> the defenders.	Bafemedi <i>M</i> the defenders.
Bahentse (the correct spelling is <i>Bafentse</i> from the verb <i>fenya</i> "win; conquer") <i>F</i> they have conquered.	Bafentse <i>M</i> they have conquered, won.
Bahenyi (the correct spelling is <i>Bafenyi</i> from the verb <i>fenya</i> "win; conquer") <i>M</i> winners, victors, conquerors.	Bafenyi <i>M</i> winners; victors.
Henyang (the correct spelling is <i>Fenyang</i> from the verb <i>fenya</i> "win; conquer") 1. <i>M</i> win 2. Defeat.	Fenyang 1. <i>M</i> win 2. Defeat.
Hetanang (the correct spelling is <i>Fetanang</i> from the verb <i>feta</i> "pass; exceed") <i>M</i> pass each other.	Fetanang <i>M</i> pass each other.
Hetang (the correct spelling is <i>Fetang</i> from the verb <i>feta</i> "pass") <i>M</i> move on, pass.	Fetang <i>M</i> move on, pass.
Hetolang (the correct spelling is <i>Fetolang</i> from the verb <i>fetola</i> "change") 1. <i>M</i> reply, respond 2. Change, transform.	Fetolang 1. <i>M</i> reply, respond 2. Change, transform.

Table 8: Names that incorrectly use [h] instead of [f]

The dictionary entry indicates that the wrong spelling entry is cross-referenced to the correct spelling entry though the correct spelling is not cross-referenced to the incorrect spelling entry. The correct spelling entry could be improved by cross-referencing the wrong spelling entry to make the user aware of the misspelt version in the following manner:

Fetolang 1. M reply, respond 2. Change, transform.
(usually misspelt as *Hetolang*).

In many old Setswana names, instead of choosing either [h] or [f], name givers have used both consonants spelt [fh]. There are therefore many names which use [fh] instead of [f], such as *Gadifhele* instead of *Gadifele*, *Kefhilwe* instead of *Kefilwe*, *Molefhe* instead of *Molefe*, *Molefhi* instead of *Molefi*, *Nonofho* instead of *Nonofo*, *Baikalafhi* instead of *Baikalafi*, and *Merafhe* instead of *Merafe*. Table 9 further illustrates this:

Wrong spelling entry	Correct spelling entry
Gadilefhe (the correct spelling is <i>Gadifele</i> from the verb <i>lefa</i> "pay a charge") M they do not pay a fine or a charge.	Gadilefe M they do not pay a fine or a charge.
Merafhe (the correct spelling is <i>Merafe</i>) 1. M tribes, ethnic groups 2. Communities 3. Nations.	Merafe 1. M tribes, ethnic groups 2. Communities 3. Nations.
Nonofho (the correct spelling is <i>Nonofo</i>) M strength.	Nonofo M strength.
Molefhi (the correct spelling is <i>Molefi</i>) M one that pays damages.	Molefi M one that pays damages.
Kehilwe (the correct spelling is <i>Kefilwe</i>) F I have been given a gift	Kefilwe F I have been given a gift.

Table 9: Names that use [fh] instead of [f]

Standard Setswana orthography does not have /fh/. The table indicates that the wrong spelling entry is cross-referenced to the correct spelling entry although the correct spelling is not cross referenced to the wrong spelling entry. The correct spelling entry could be improved by cross-referencing the wrong spelling entry to make the user aware of the misspelt version in the following manner:

Merafe 1. M tribes, ethnic groups 2. Communities 3. Nations.
(usually misspelt as *Merafhe*).

4.1 The use of the [ny] and [n] variants

There is confusion amongst Setswana speakers between words that end with the causative [ny-] and the reciprocal [an-]. For instance, in daily speech there is an enduring confusion between *puisano* and *puisanyo* where *puisano* means a discussion or dialogue and *puisanyo* strictly speaking is an act of making each other talk. This confusion is reflected in many names which should end in [-n-] but end in [-ny-]. There is therefore *Kedibonye* instead of *Kedibone*, *Golekanye* instead of *Golekane*, *Kebabonye* instead of *Kebabone*, *Kemmony*, instead of *Kemmone*, *Gareitsanye* instead of *Gareitsane*, *Kelekanye* instead of *Kelekane* and *Lesanyang* instead of *Lesanang*. The dictionary captures both spelling variants in its macro-structure as shown in the sampled entries in Table 10.

Wrong spelling entry	Correct spelling entry
Kedibonye (the correct spelling is <i>Kedibone</i>) <i>F</i> I saw them, or I have seen them; I have been a victim of circumstances.	Kedibone <i>F</i> I saw them, or I have seen them; I have been a victim of circumstances.
Kemmony (the correct spelling is <i>Kemmone</i>) 1. <i>F</i> I have seen him or her 2. I have found him or her.	Kemmone <i>F</i> 1. I have seen him or her 2. I have found him or her.
Golekanye (the correct spelling is <i>Golekane</i>) <i>F</i> it is enough; it is sufficient.	Golekane <i>F</i> it is enough; it is sufficient.
Gareitsanye (the correct spelling is <i>Gareitsane</i>) <i>M</i> we do not know each other.	Gareitsane <i>M</i> we do not know each other.

Table 10: Names that incorrectly use [ny-] instead of [n-]

The wrong spelling entry is cross-referenced to the correct spelling entry even though the correct spelling is not cross referenced to the wrong spelling entry. The correct spelling entry could be improved by cross-referencing the wrong spelling entry to make the user aware of the misspelt version in the following manner:

Kedibone *F* I saw them, or I have seen them; I have been a victim of circumstances (usually misspelt as *Kedibonye*).

4.2 Names which should have a double [n], or double [m] but lack this feature

In Setswana orthography, the long nasal sound is represented orthographically by doubling the concerned nasal consonant, whether it is /n/ or /m/, while a short nasal is written as a single nasal consonant. Doubling the long nasal in orthography is supported by the fact that one of the nasal consonants is a tone

bearing syllabic nasal. Spelling variation is, however, introduced when there is an inconsistency in the writing of a long nasal. Some Setswana names that are pronounced with a long nasal are spelt with a single letter as if they use a short nasal which is represented by a single character. This has caused variation in the spelling of names, with some spelt with a double nasal while others have a single nasal character. The dictionary therefore has *Gaofengwe* /χáúfɪŋwí/ as well as the correctly spelt *Gaofenngwe* /χáúfɪŋwí/, and *Mamosadinyana* /mámósádɪjáná/ as well as the correctly spelt *Mmamosadinyana* /mmámósádɪjáná/ as shown in Table 11.

Wrong spelling entry	Correct spelling entry
Gaseikangwe (the correct spelling is <i>Gaseikanngwe</i>) <i>F</i> it is not trusted or relied on.	Gaseikanngwe <i>F</i> it is not trusted or relied on.
Molotsi (the correct spelling is <i>Mmolotsi</i>) <i>M</i> one who takes animals from the kraal.	Mmolotsi <i>M</i> one who takes animals from the kraal.
Mamosadinyana <i>F</i> (the correct spelling is <i>Mmamosadinyana</i>) queen or Queen of England.	Mmamosadinyana <i>F</i> the queen of England.

Table 11: Names that incorrectly miss a pre-nasal nasal consonant

The wrong spelling entry is cross-referenced to the correct spelling entry though the correct spelling is not cross-referenced to the wrong spelling entry. The correct spelling entry could be improved by cross-referencing the wrong spelling entry to make the user aware of the misspelt version in the following manner:

Mmolotsi *M* one who takes animals from the kraal. (usually misspelt as *Molotsi*).

5. The impact of Northern Setswana on name spellings

Some spelling variations in the dictionary were brought about by the influence of Northern Setswana. Northern Setswana is a cluster of dialects spoken by Bangwato, Batawana, and Bakhurutshe, amongst other groups who are found in the central, northern and parts of north-eastern Botswana (Cole 1955). Northern Setswana is sometimes referred to as Sengwato in the literature (Tlale 2005, Boyer and Zsiga 2013).

5.1 The loss of /t/ and /tʰ/

Historically, Northern Setswana uses the voiceless alveolar plosive /t/ where the

standard language uses the voiceless unaspirated affricate /tʃ/ and usually replaces the voiceless aspirated affricate /tʃh/ consonantal sound (Batibo 1999) with the voiceless aspirated explosive /tʃh/. This is reflected in many words in the Setswana vocabulary as in Table 12 that illustrates the different orthographic representations of some words between Northern Setswana and other Setswana dialects.

Standard Setswana orthography	Northern Setswana	English
tlhaga	Thaga	Grass
otlhaya	othaya	Discipline
setlha	Setha	Bladder
tlhaba	Thaba	stab, pierce
tla	Ta	Come

Table 12: Standard Setswana vs Northern Setswana dialect

This evolution has had a lasting influence on how speakers of the Northern Setswana dialect spell personal names. The spelling of the personal names, therefore, follows their pronunciation and not the Setswana standard orthography. The dictionary has included spellings of standard Setswana and that of Northern Setswana leading to a variation of names in which the same name is spelt with /tʃh/ and /tʃ/. For instance, *Tlhalefang* and *Thalefang*, *Atlholang* and *Atholang*, *Baikatlhodi* and *Baikathodi*, *Baiphethli* and *Baiphethi*, *Batlang* and *Batang*, *Batlhaetse* and *Bathaetse*, and *Batlhalefi* and *Bathalefi*. The Northern Setswana name spellings are captured in the dictionary as they are so spelt by their owners. This is in line with the recommendations of Setswana Standard Orthography of 1981 which notes that, "As a general rule, proper names of people should be spelled as used by the persons themselves" (p.19). The dictionary, however, cross references the standard spelling entry at each Northern Setswana spelling entry, for the benefit of the user. This is demonstrated in Table 13.

Northern Setswana spelling entry	Standard Setswana spelling entry
Atholang (Northern Setswana version of <i>Atlholang</i>) <i>M</i> pass/give judgement (pl.).	Atlholang <i>M</i> pass/give judgement (pl.).
Baikathodi (Northern Setswana version of <i>Baikatlhodi</i>) <i>M</i> those who judge themselves; those who find themselves guilty.	Baikatlhodi <i>M</i> those who judge themselves; those who find themselves guilty.
Molothanyi (Northern Setswana version of <i>Molotlhanyi</i>) <i>M</i> one who causes conflicts between others	Molotlhanyi <i>M</i> one who causes conflicts between others.

Mothagodi (Northern Setswana version of <i>Motlhagodi</i>) <i>M</i> one who cultivates; one who clears grass.	Motlhagodi <i>M</i> one who cultivates; one who clears grass.
---	--

Table 13: Northern Setswana names

5.2 Conflation of prefix *le-* and *lo-*

Northern Setswana also does not make a distinction between the noun class prefix /*lo-*/ of class 11 and the noun class prefix /*le-*/ of class 5. The distinctions have been collapsed into singular prefix [*le-*] of class 5. This difference has been discussed by Cole (1955). All the words which in other Setswana dialects belong to class 11 noun class, in Northern Setswana they are found in class 5. According to general Setswana books (e.g. Mogapi 1984; Cole 1955) and Setswana Standard Orthography, the determination of whether a word takes a [*le-*] or [*lo-*] prefix is based on the plural prefix that the word takes (Chebanne et al. 2008). When class 11 nouns are pluralised, they take class 8 prefix [*di-*] and when class 5 nouns are pluralised, they take the class 6 prefix [*ma-*] (Cole 1955; Chebanne et al. 2008). This may be summarised as follows:

Singular noun class prefix	→	Plural noun class prefix
noun class 5 [<i>le-</i>]	→	noun class 6 [<i>ma-</i>]
noun class 11 [<i>lo-</i>]	→	noun class 8 [<i>di-</i>]

Other Setswana dialects	Northern Setswana	English
logong	Legong	Wood
logopo	Legopo	Rib
loleme	Leleme	tongue
lobadi	Lebadi	scar
lofuka	Lehuka	feather

Table 14: [*lo-*] and [*le-*] noun variation

These /*le-*/ and /*lo-*/ variations have been extended to personal names, and as result we have the same name with different orthographic representations, for example, *Leungo/Loungo*, *Lerato/Lorato*, *Losika/Lesika*, *Lonyatso/Lenyatso*, *Logong/Legong* and *Loapi/Lewapi*. If we are to apply the pluralisation rule to determine the correct spellings of the names, then the variations of *Lotshwao*, *Loeto*, *Loungo* and *Legong* will be incorrect. This is because their pluralisation would flout the standard pluralisation rule. The first three names would take the noun class 8 prefix [*di-*] to become *ditshwao*, *dieto* and *diungo*, while the last word will take

the noun class 6 prefix [ma-] and become *magong*, all of which are linguistically non-existent in Setswana vocabulary. Table 15 further illustrates this as captured in the dictionary:

Correct spelling entry	Wrong spelling entry
Lebopo (also <i>Lobopo</i>) <i>M</i> the universe; creation.	Lobopo (also <i>Lebopo</i>) <i>M</i> creation; the universe.
Letshwao 1. <i>M</i> a sign 2. a symbol 3. a brand.	Lotshwao (the correct spelling is <i>Letshwao</i>) <i>M</i> a mark or sign.
Leungo <i>F</i> a fruit	Loungo (the correct spelling is <i>Leungo</i>) <i>F</i> a fruit
Leeto 1. <i>M</i> a visit 2. a journey.	Loeto (the correct spelling is <i>Leeto</i>) <i>M</i> a visit; a journey.
Logong <i>M</i> a piece of wood.	Legong (the correct spelling is <i>Logong</i>) <i>M</i> a piece of wood.

Table 15: [le] and [lo-] Setswana names

As previously stated, the Northern Setswana name spellings are captured in the dictionary as they are spelt by their owners, as recommended by the *Standard Setswana Orthography of 1981*, however, both names would be captured twice, one being the correct version.

The wrong spelling entry is cross-referenced to the correct spelling entry though the correct spelling is not cross-referenced to the wrong spelling entry. The correct spelling entry could be improved by cross-referencing the wrong spelling entry to make the user aware of the misspelt version in the following manner:

Leeto 1. *M* a visit 2. a journey. (usually misspelt as *Loeto*).

6. Discontinued consonants in Setswana names

An orthographic decision was taken to discontinue the use of certain consonants in the orthography of Setswana. A conference held in Johannesburg, South Africa on April 25, 1937 attended by representatives of the Education Departments of Bechuanaland Protectorate, the Cape, the Orange Free State, and the Transvaal, together with the Inter-University Committee for African Studies, resolved to change how it spelt two consonants: the voiceless postalveolar affricate [tʃʰ] and the voiceless postalveolar fricative /ʃ/. The conference resolved that [tʃʰ] which was then spelt as [ch], should henceforth be spelt as [tʃh], while /ʃ/ which was then spelt as [sh] was to be spelt as [ʃ]. It was noted that:

The symbol *š*, with its movable diacritic, was chosen by the Committee to represent the voiceless hushing fricative in preference of the I.P.A. symbol on the principle of avoiding new types, and to the symbol *sh* on account of the latter's ambiguousness and the difficulties it creates in connection with the representation of the hissing and hushing fricative combinations. It was felt that the symbol *š* possessed all the advantages of the I.P.A. symbol, without its disadvantages as a new type, and with the added advantage that its movable diacritic indicates, without exaggerating, the dialectal differences between various Tswana dialects among themselves and between Tswana and Transvaal Sotho and Basutoland Sotho in general. (Lestrade 1937: 8-9)

While this decision was taken over 80 years ago and has been adopted in much of Setswana writing patterns, it has had no impact on the spelling of Setswana personal names. Batswana still use [ch] and [sh] in the spelling of their names. No name in the Setswana database that informed the writing of *A Dictionary of Setswana Personal Names* used [tšh] or [š]. This may be because most Batswana are bilingual in Setswana and English which uses the [ch] and [sh] characters and therefore associate the [tʃ^h] and /ʃ/ sounds with the English [ch] and [sh] representations. The other reason may be related to advances in technology — that [š] is a more challenging character to write on a typewriter or computer. Consequently, [ch] and [sh] have endured in Setswana names. Names that use [ch] or [sh] are therefore entered in the dictionary as they are spelt by their owners. They are not spelt using the agreed post 1937 orthography. However, there is a note in the microstructure of the dictionary on the correct spelling as demonstrated in the examples below:

- Chaba** *M* (an old spelling of *Tšhaba*); nation.
- Chabadiile** *M* (an old spelling of *Tšhabadiile*) nations have gone.
- Chabaesele** *M* (an old spelling of *Tšhabaesele*) a different nation.
- Chabaditsile** *M* (an old spelling of *Tšhabaditsile*) nations have arrived.
- Chabaefe** *M* (an old spelling of *Tšhabaefe*) which nation?
- Cheleketo** *M* (old spelling of *Tšheleketlo*); a struggle; suffering.

Examples from the dictionary of names that use the discontinued [sh] instead of the standard orthography character [š] are shown as follows:

- Bashapi** (an old spelling of *Bašapi*) **1.** *M* swimmers **2.** those who beat.
- Bashi** (an old spelling of *Baši*) *M* short for *Basimane*; boys. A pet name for a boy from *Bashimane*.
- Ramasheleng** (the correct spelling is *Ramašeleng*) **1.** *M* the father of shillings **2.** the man of shillings
- Boshafatso** (an old spelling of *Bošafatso*) *M* renewal.
- Boshula** (an old spelling of *Bošula* — the standard spelling of this word is *Bosula*) **1.** *M* bad taste **2.** evil
- Shianyana** (the correct spelling is *Šianyana*) *M* small boy.

This note on the correct spelling must not be misconstrued as a cross-reference. It does not point to another entry elsewhere in the dictionary. It merely states how an entry would be written in Setswana standard orthography. Entries with discontinued characters [sh] and [ch] are therefore entered in the dictionary exactly the way the owners of the names spell them.

7. Conclusion

This paper has examined spelling variations in *A Dictionary of Setswana Personal Names*, highlighting the phonological and morphological patterns that underlie these differences. It has shown that many of these variations stem from the influence of distinct Setswana dialects — particularly the northern and southern varieties — on the orthographic representation of personal names. While general dictionaries tend to be normative, adhering to standardized orthographies, this study reveals that formal spelling rules are sometimes bypassed in favor of forms that more closely reflect speakers' individual pronunciations.

Despite the introduction of the phoneme /š/ into Setswana orthography over eighty years ago, this study found no personal names utilizing this symbol. Instead, speakers continue to use [ch] and [sh], which are considered outdated under current orthographic standards. *A Dictionary of Setswana Personal Names* documents these spelling variants as provided by name givers or bearers, aligning with the 1981 Standard Setswana Orthography's recommendation that personal names be recorded as they are used by individuals.

Ultimately, this paper underscores the ubiquity of variation in natural languages and emphasizes the need for lexicographic strategies that enhance user accessibility. One such strategy proposed here is the documentation of personal names in their authentic forms, supplemented by cross-referencing and informational annotations to guide users through variant spellings.

References

Dictionaries

- Kgasa, M. and J. Tsonope. 1995. *Thanodi ya Setswana*. Gaborone: Longman.
- Otlogetswe, T.J. 2012. *Tlhalosi ya Medi ya Setswana*. Gaborone: Medi Publishing.
- Otlogetswe, T.J. and G.N. Ramaeba. 2024. *A Dictionary of Setswana Personal Names*. Gaborone: Mokala Publishing.

Online Sources

- African Pidgin Dictionary*.
www.pegeen.com
- Spelling Variants: Why Are They Important?*
<https://blog.spellquiz.com/spelling-variants>

Academic papers and books

- Batibo, H.M.** 1999. The Grammaticalization Process in Setswana. *Afrikanistische Arbeitspapiere (AAP)* 59: 131-150.
- Batibo, H.M.** 2013. The Evolution and Adaptation of Swahili and Tswana Syllable Structures. Legère, K. (Ed.). 2013. *Bantu Languages and Linguistics: Papers in Memory of Dr Rugatiri D.K. Mekacha*. Bayreuth African Studies 91: 13-34. Bayreuth: Bayreuth University Press.
- Baugh, J.** 2009. Linguistic Diversity, Access, and Risk. *Review of Research in Education* 33(1): 272-282.
- Boyer, O. and E. Zsiga.** 2013. Phonological Devoicing and Phonetic Voicing in Setswana. Orié, Olanike Ola and Karen W. Sanders (Eds.). 2013. *Selected Proceedings of the 43rd Annual Conference on African Linguistics (ACAL): Linguistic Interfaces in African Languages, New Orleans, USA, 15-17 March 2012*: 43: 82-89. Somerville, MA: Cascadilla.
- Chabata, E.** 2003. Linguistic Variation in Shona with Special Reference to Monolingual Dictionaries. *Lexikos* 13: 81-91.
- Chebanne, A., T.J. Otlogetswe, P. Seloma, A. Molosiwa, B. Moatlhodi, N. Kgolo and M. David.** 2008. *Mokwalo o o lolameng wa Setswana*. Gaborone: Pentagon Publishers.
- Cole, D.T.** 1955. *An Introduction to Tswana Grammar*. Cape Town: Longman.
- DALL (Department of African Languages and Literature), University of Botswana.** 1999. *The Sound System of Setswana*. Gaborone: Lightbooks.
- Lestrade, G.P.** 1937. *The Practical Orthography of Tswana — Recommendations of the Tswana District Orthography Committee. Approved by the Union Advisory Committee on African Studies and Research*.
- Lichtenstein, H.** 1928. *Travels in Southern Africa in the Years 1803, 1804, 1805 and 1806*. (A reprint of the translation from the original German by Anne Plumptre). Translated by Anne Plumptre. Cape Town: The Van Riebeeck Society.
- Lichtenstein, H.** 1973. *Foundation of the Cape (1811) and About the Bechuanas (1807)*. Edited and translated by O.H Spohr. Cape Town: A.A. Balkema.
- Livingstone, D.** 1875–1885. *Missionary Travels and Researches in South Africa*. London: J. Murray.
- Mareme, G.B.** 2007. *Thanodi ya Setswana*. Pietermaritzburg: Nutrend Publishers.
- Ministry of Education.** 1981. *Setswana Standard Orthography of 1981*. Gaborone: Ministry of Education, Government of Botswana.
- Mogapi, K.** 1984. *Thutapuo ya Setswana*. Gaborone: Longman.
- Motzafi-Haller, P., K. Keuthmann and R. Vossen.** 2006. Setswana Dialects and Inter-dialectal Variation in the Republic of Botswana. I. Setswapong. *Zeitschrift der Deutschen Morgenländischen Gesellschaft* 156(1): 9-41.
- Mulingi, Y.** 2023. *Development of Models for Non-word Error Detection for and Correction System for Kiswahili Language*. Master's dissertation, Kampala International University, Uganda.
- Murray, R.W.** 1996. Historical Linguistics: The Study of Language Change. O'Grady, W., M. Dobrovolsky and F. Katamba. 1996. *Contemporary Linguistics. An Introduction*: 313-371. London/New York: Longman.
- Otlogetswe, T.J.** 2014. Extracting Business Terms for Dictionary Subject Label. Khumalo, L. (Ed.). 2014. *African Languages and Linguistic Theory: A Festschrift in Honour of Professor Herbert Chimhundu*: 268-286. Cape Town: CASAS.
- Otlogetswe, T.J.** 2015. Treatment of Spelling Variants in Setswana Monolingual Dictionaries. *Lexikos* 25: 262-284.

- Prah, K.** 1998. *Between Distinction and Extinction: The Harmonization and Standardization of African Languages*. Cape Town: CASAS.
- Prinsloo, D. and N.D. van Graan.** 2021. Principles and Practice of Cross-referencing in Paper and Electronic Dictionaries with Specific Reference to African Languages. *Lexicography, Journal of ASIALEX* 8(1): 32-58.
- Scontras, G., Z. Fuchs and M. Polinsky.** 2015. Heritage Language and Linguistic Theory. *Frontiers in Psychology* 6: 1545.
- Sebina, B. and T.J. Otlogetswe.** 2023a. Vowel Sex-based Symbolism in Setswana Personal Names. *Studies in African Linguistics* 52(13): 30-44.
- Sebina, B. and T.J. Otlogetswe.** 2023b. Syllable Structure in Setswana Personal Names. *Voprosy onomastiki* 20(2): 174-192.
- Tlale, O.** 2005. *The Phonetics and Phonology of Sengwato, a Dialect of Setswana*. Unpublished PhD thesis, Georgetown University, Washington, DC.

Lexicographic Post-processing and Refinement Issues in Wordnets: The Ukrainian Wordnet Experience

Maksym Vakulenko, *Darmstadt University of Applied Sciences,
Darmstadt, Germany*
(maxvakul@yahoo.com)
(<https://orcid.org/0000-0003-0772-7950>)

Abstract: This paper considers challenges arising within the development of the Ukrainian wordnet, specifically focusing on the integration and alignment of vocabularies originating from diverse sources, such as conventional dictionaries, the Open English Wordnet, and other wordnets. Various cases of synonymy and hyper-hyponymy presented in the Open English Wordnet and in conventional terminological and phonetic dictionaries were analyzed, as well as the definitions of terms in these sources were compared. It was found that traditional dictionaries do not contain all the necessary hyper- and hyponymic relations, and definitions in wordnets are often inaccurate or incomplete. It is shown that creating wordnets based on the "merge" and combined approaches, exemplified by Ukrajinet, necessitates specific lexicographic post-processing. This procedure involves several crucial steps, including incorporating additional lexical units for lacking concepts, refining definitions, and introducing novel concepts and semantic structures, particularly those associated with hyper- and hyponymy, aligned with revised definitions.

Keywords: WORDNETS, UKRAJINET, CONCEPT, DEFINITION, LEXICAL-SEMANTIC RELATIONS, SYNONYMY, HYPER-HYPONYMY, LEXICOGRAPHIC DATA

Opsomming: Leksikografiese postverwerking en verfyningskwessies in woordnette: Die Oekraïense woordnet-ervaring. Hierdie bydrae beskou uitdagings wat binne die ontwikkeling van die Oekraïense woordnet (Ukrajinet) ontstaan, en fokus spesifiek op die integrasie en belyning van woordeskat wat afkomstig is van diverse bronne, soos konvensionele woordeboeke, die Open English Wordnet en ander woordnette. Verskeie gevalle van sinonimie en hipero-hiponimie wat in die Open English Wordnet en in konvensionele terminologiese en fonetiese woordeboeke aangebied word, is geanaliseer. Die definisies van terme in hierdie bronne is ook met mekaar vergelyk. Daar is bevind dat tradisionele woordeboeke nie al die nodige hipero- en hiponimiese verwantskappe bevat nie, en definisies in woordnette is dikwels onakkuraat of onvolledig. Daar word aangetoon dat die skep van woordnette gebaseer op die "samesmeltings-" en gekombineerde benaderings, soos geïllustreer deur Ukrajinet, spesifieke leksikografiese postverwerking noodsaak. Hierdie prosedure behels verskeie belangrike stappe, insluitend die inkorporering van bykomende leksikale eenhede vir ontbrekende konsepte, die verfyning van definisies en die bywerking van nuwe konsepte en semantiese strukture, veral dié wat met hipero- en hiponimie geassosieer word sodat dit met hersiene definisies strook.

Sleutelwoorde: WOORDNETTE, UKRAJINET, KONSEP, DEFINISIE, LEKSIKAAL-SEMANTIESE VERHOUDINGS, SINONIMIE, HIPERO-HIPONIMIE, LEKSIKOGRAFIESE DATA

1. Introduction

The rapid development of linguistic resources such as wordnets (Fellbaum 1998; Bond et al. 2016; McCrae et al. 2020; Siegel and Bond 2021) caused a breakthrough in lexicography, providing multilingual lexical databases containing definitions, use examples, lexical-semantic relations (LSRs), etc. Wordnets are useful in various natural language processing tasks, such as machine translation and semantic text analysis. In general, there are two approaches to developing wordnets: "expansion" and "merge", where the first is based on translations of existing lemma forms and the second foresees building semantic structures of the language in question (synonymy, hyper-/hyponymy, mero-/holonymy, and antonymy) with subsequent linking of concepts through interlingual indices — ILIs (Bond et al. 2016: 50-51). Then the development of wordnets based on the "merge" and combined ("expansion" and "merge") approaches necessitates the joined use of lexicographic information available in existing wordnets, mainly Open English Wordnet (OEWN)¹, and traditional dictionaries containing lexical units and their definitions.

In particular, the first version of the Ukrainian wordnet, Ukrajinet 1.0², was created based on the merge approach, making use of the "Explanatory Dictionary on Physics" (TSF)³ and the "Dictionary of Physical Synonymous Terms" (SFS)⁴. As the follow-up of this project, the 750 concept definitions and 700 examples of use have been added so far, and this resource was supplemented with a translation of OEWN into Ukrainian. The definitions were taken from the "Dictionary of Ukrainian Language in 20 volumes" (SUM)⁵, TSF, and OEWN. The current version of Ukrajinet contains 153 683 lexical entries (single and multi-word lexemes) combined in 124 214 synsets.

The linguistic data coming from different sources pose "a particular challenge for annotators" (Rudnicka et al. 2019: 310), requiring tedious refinement and adjustment. Extremely difficult is to find corresponding concepts in different languages where some of them may be absent, and others may differ in their scope that requires thorough analyzing definitions and use examples. For example, the concept i29195 "put" — 'cause to be in a certain state; cause to be in a certain relation' — corresponds to several Ukrainian concepts, in particular: 1 *викликати [в когось настрій/стан]* 'put [someone in a mood/state]'; 2 *оформляти* 'put [something] in form'; 3 *упорядковувати* 'put [something] in order'.

The most problematic are here the lexical units that enter the corresponding LSRs, and their definitions. For linking wordnets, the most important relations are synonymy and hypernymy, where the greatest number of equivalents across synsets comes from absolute synonymy (Rudnicka et al. 2019). So, synonymy plays a fundamental role in creating wordnets. It is remarkable that if synonymy is defined on the basis of hypernymy, treating hypernymy as the fundamental relation, the resulting synsets would differ (Halas Popović 2023).

Synonymy is widespread in common language and also in various terminologies (Vakulenko 2018; Bokshanj 2021). However, this concept is treated differently in different sources. For example, the Longman dictionary⁶ defines a synonym as "a word with the same meaning as another word in the same language" (Longman: 1685), where only lexical synonymy is considered and no partial synonymy and cross-language synonymy are recognized.

In contrast, Löbner understands synonymy in a way that:

"Two expressions are synonymous if they have the same meaning" (Löbner 2013: 203). This is an interesting approach where the synonymy is taken more generally, involving expressions instead of a single word and depending on their meanings.

The more general definition is based on the notion of a concept and suggests that: "Synonyms are the units of a language system (morphemes, lexemes, syntactic constructions) that within usually one linguistic level represent the same concept, and that can be substituted in the context without distorting the general sense" (Vakulenko 2018: 36). A similar approach is adopted for wordnets (Fellbaum 2006; Rudnicka et al. 2019: 299) where synonymy is considered in a narrower sense as restricted to the lexical level.

Another approach to synonymy that is aimed at meeting the lexicographer's requirements, derives synonymy from the so-called constitutive relations such as hypernymy/hyponymy, meronymy/holonymy, and antonymy (Maziarz et al. 2013; see also Rudnicka et al. 2019: 300).

The theory of synonymy in the context of terminological research was further developed in Vakulenko (2018: 40-43) where it was proposed to discern formal synonymy (covering lexical and grammatical synonymy) and contentual one. In particular, terminological dictionaries in the fields of physics (WPEDFR⁷; WPED⁸; TSF; SFS) and phonetics (Yurtbaşı 2017a; 2017b; Vakulenko 2018: 49-55) contain various kinds of full synonyms resulting from these types of synonymy. In what follows, these notions will be used.

A large problem in developing wordnets, particularly Ukrajinet, is the coordination and adjustment of lexicographic data coming from different sources. This problem will be addressed here. Synonymy and hyper-/hyponymy in wordnets will be considered and traditional dictionaries in connection with the existence of "hidden" senses not explicitly displayed in the last sources. A comparative analysis will be provided of definitions available in OEWN in comparison with those coming from academic dictionaries.

The article is organized as follows. In Section 2, we describe the method and material used for this research. Section 3 is devoted to the study of synonymy and hypernymy/hyponymy as they appear in traditional terminological dictionaries. In Section 4, we carry out a comparative analysis of term definitions available in OEWN and traditional dictionaries. Section 5 provides a discussion of related issues and future plans. The concluding remarks are presented in Section 6.

2. Methods and material

The statistical (in the symptomatic sense) and analytical methods of terminology science (see Vakulenko 2018) are applied to carry out the comparative analysis of terminological units and their definitions coming from conventional dictionaries in physics and phonetics (WPEDFR; WPED; TSF; Yurtbaşı 2017a; 2017b), with those of OEWN. The statistical method determines in this case the availability of terms and their definitions, as well as the use examples, while the analytical method reveals scientific quality of these. The study is restricted to English, German, and Ukrainian lexical units, and uses standard conventions for traditional dictionaries. Within this lexicographic treatment, the units appearing in the same terminological article, are considered full synonyms, and the nouns in the word combinations which are specified by preceding adjectives (modifiers), are considered hypernyms. For example, *shift* is a hypernym for *Doppler shift*. As for definitions, the term to which the described term is referred is regarded as its hypernym.

3. Synonymy and hypernymy

Approximately 10% of concepts available in the physical dictionaries (WPEDFR; WPED; TSF) have synonyms. The rise of lexical synonymy is conditioned by a certain degree of semantic similarity of concepts whose hyponyms have identical meanings:

- Doppler **shift** – Doppler **displacement** – Doppler **frequency shift** (WPEDFR: 375);
- Ger. Polschuh**abstand** m – Luftspalts**länge** f – Luftspalts**breite** f (WPEDFR: 570);
- **initiation** of the discharge – **ignition** of the discharge (WPEDFR: 726);
- level of significance – significance **level** – error **probability** (WPEDFR: 838);
- **level** of significance – confidence **level** – **degree** of confidence – confidence **probability** – confidence **coefficient** (WPEDFR: 838);
- Ger. statistische Sicherheit – Sicherheitss**chwelle** f – Sicherheits**grad** m – Vertrauens**koeffizient** m – Konfidenz**koeffizient** m – Konfidenz**niveau** n – Konfidenz**wahrscheinlichkeit** f – Bedeutsamkeit**stufe** f (WPEDFR: 838);
- contact **clearance** – contact **gap** – contact **spacing** (WPED: 594);
- **monitor** – monitoring instrument – survey **instrument** – survey **meter** (WPED: 594);
- Ger. spiegelnder **Körper** – weisses **Objekt** (WPED: 306);
- sighting **apparatus** – sighting **device** (WPED: 594);
- Ger. Visiervor**richtung** f – Zielgerät n (WPED: 594);
- **density** – *specific weight*, optical **isomer** – optical **antipode** (TSF: 165);
- *light-beam oscillograph* – *light-beam instrument* (TSF: 165);
- **argument** – *independent variable*, **momentum** – *quantity of motion* (Vakulenko 2018: 41);

- *magnetometer* – *magnetic(-field) meter* – *magnetic detector* – *gaussmeter* (TSF: 296-297);
- *approximant* – *frictionless continuant* (Vakulenko 2018: 53);
- *diminuendo accent* – *falling accent* – *diminuendo stress* – *falling stress* (Vakulenko 2018: 53).

In these examples, there is a clear semantic correlation between the head terms, including hypernymy/hyponymy (*shift* – *frequency shift*, *instrument* – *monitor*, *instrument* – *meter*, *weight* – *density*, *instrument* – *oscillograph*, *variable* – *argument*, *quantity* – *momentum*, *meter* – *detector*, *continuant* – *approximant*) and synonymy (*shift* – *displacement*, *level* – *degree*, *accent* – *stress*). It can be seen from this that a necessary condition for the synonymy of two lexemes is the presence of hyponyms in the corresponding concepts that are part of the same synset. So, the equivalence of hyponyms is caused by synonymy of their hypernyms. As can be seen from the synonymous pair *Doppler shift* – *Doppler frequency shift*, the opposite is correct only when the hypernym (*shift*) has an additional meaning identical to the meaning of its hyponym (*frequency shift*).

Grammatical variability also often causes synonymy, primarily for compound terms. This is exemplified by the following series of synonyms: *unshared electron pair* – *electron lone pair* – *lone pair* – *lone-pair electrons* (TSF: 394). Here the element *pair* that is a hypernym to *electron lone pair*, can be considered as a hypernym to the unit *electrons* that is a hypernym to *lone-pair electrons*.

The situation is somewhat different for the **contentual synonymy** that arises when different terms reflect different aspects of the concept or different approaches to its study (Vakulenko 2018: 41). This kind of synonymy is also typical for composed terms, i.e. when some senses of their hypernyms are neutralized. In traditional dictionaries, the corresponding head terms for synset members often display either little semantic similarity or complicated and unstraightforward semantic relations:

- piezo effect* – *piezoelectric effect* – *polarization* of distortion (WPED: 594);
- Ger. *schwarzer Körper* – *schwarzer Strahler* 'blackbody radiator' (WPED: 22);
- density matrix* – *statistical matrix* – *statistical operator* (TSF: 305);
- Gibbs energy* – *Gibbs thermodynamic potential* – *isobaric-isothermal potential* – *free enthalpy* (TSF: 166);
- thermal electron emission* – *Richardson effect* – *filament emission* (TSF: 165);
- Planck constant* – *quantum* of action (TSF: 608);
- isotopic indicator* – *tracer* – *tracer isotope* – *label* – *label(l)ed atom* (TSF: 225);
- freezing point* – *freezing temperature*; *dielectric loss coefficient* – *dielectric loss factor* – *dielectric loss index* – *dielectric power factor* – *dielectric dissipation factor* – *dielectric loss tangent* (Vakulenko 2018: 43).

Historically, these terms entered the conventional dictionaries due to their practical use that from the standpoint of terminological theory is the result of the statistical method in the symptomatic sense (Vakulenko 2018: 14-19). However, the

appearance of synonymy between composed terms assumes synonymy or hypernymy between head terms describing corresponding concepts. In other words, the conceptual approach underlying the development of wordnets necessitates the introduction of extra meanings of these head terms that have not been explicitly displayed in conventional dictionaries. Such augmentation of the human lexicographic practice manifests the application of the analytical method of terminology science (Vakulenko 2018: 14-19).

While the relative number of such correlated and under-refined concepts is an open question, their existence was already observed while working directly on Ukrajinet.

In particular, the synset **omw-1354-n** contains lexemes "НВЧ-генератор" 'microwave oscillator' and "лампа зворотної хвилі" 'backward-wave tube' which, according to the TSF, belong to the nests of different terms: "генератор" 'oscillator' (**omw-124123-n**) and "лампа" 'tube' (**omw-1617-n**), respectively. The synset **omw-689-n** (i61958) contains lexemes "гомоморфізм" 'homomorphism' and "відображення гомоморфне" 'homomorphic mapping' which, according to the TSF, are also included in different term nests: "подібність" 'similarity' (**omw-100-n**) and "відображення" 'mapping' (**omw-54804-n**). In such cases, lacking senses are added to the corresponding nesting terms.

4. Dictionary definitions vs OEWN definitions

The degree of semantic similarity between hypernyms to the abovementioned synsets cannot be easily calculated, as it strongly depends on the corresponding definitions (see Vakulenko 2019). For example, the Open English Wordnet treats energy as "a thermodynamic quantity equivalent to the capacity of a physical system to do work", connecting this concept with the hypernym "quantity" (<https://en-word.net/lemma/energy>). In turn, the TSF defines energy as "a general measure of various processes and types of interaction that allows measuring all kinds of physical forms of movement and interaction", relating this concept to the hypernym "measure". Using the latter is recommended, because some definitions of physical and technical terms available in the OEWN, fail to correctly convey significant features of the defined concepts such as those with the following ILIs.

Incorrect OEWN definitions

Determinant (i80590): 'a square matrix used to solve simultaneous equations'. This definition is completely incorrect. It is recommended to use the latter formulation from the TSF that is freely available and defines this concept as 'a polynomial, each term of which is a product of matrix elements, and only one coefficient is included in the product from each row and each column of the matrix' ['поліном, кожен член якого є добутком елементів матриці, причому з кожного рядка і кожного стовпця матриці в добуток входить лише один співмножник'].

Field, field of force, force field (i97792): 'the space around a radiating body within which its electromagnetic oscillations can exert force on another similar body not in contact with it'. This definition is not correct, because the force field is not about a body radiating electromagnetic oscillation. The Longman (which is freely available) defines this concept better as 'the area in which the natural force is felt or has an effect'. The TSF treats this concept more comprehensively, as 'a part of space (limited or unlimited), at each point of which a force with a certain numerical value and direction acts on a material particle located there, which depends only on the coordinates of this point' ['частина простору (обмежена чи необмежена), у кожній точці якої на розміщену там матеріальну частинку діє певна за числовим значенням і напрямком сила, що залежить тільки від координат цієї точки']. The last definition is recommended.

Luminescence (i97901): 'light not due to incandescence; occurs at low temperatures'. This description is not relevant to luminescence. However, the Longman also does not highlight essential properties of luminescence, defining this phenomenon as 'a soft shining light'. The TSF treats luminescence as 'radiation, which is an excess over the thermal radiation of the body and continues for a time that significantly exceeds the period of light oscillations' ['випромінювання, що являє собою надлишок над тепловим випромінюванням тіла і продовжується протягом часу, який значно перевищує період світлових коливань'] that should be used in wordnets.

Paramagnetism (i97922): 'materials like aluminum or platinum become magnetized in a magnetic field but it disappears when the field is removed'. This definition is incorrect, because paramagnetism is a property, not a material. The TSF contains a definition that should be used instead: 'the property of substances (paramagnets) to be magnetized in the direction of an external magnetic field' ['властивість речовин (парамагнетиків) намагнічуватися в напрямку зовнішнього магнітного поля'].

Vibration, oscillation (i75292): 'a regular periodic variation in value about a mean'. These terms are not synonyms, as can be seen from the Longman that proposes two senses for vibration(s): 1 'a continuous slight shaking movement'; 2 'vibes'; and four senses for oscillation(s): 1 'frequent changes between two extreme amounts of limits'; 2 'frequent changes between one feeling or attitude and another'; 3 'a regular movement of something from side to side'; 4 'a regular change in direction of an electrical current'. This treatment is much more relevant and should be applied in wordnets.

Inexact OEWN definitions

Elasticity (i63383): 'the tendency of a body to return to its original shape after it has been stretched or compressed'. This definition uses common language and lacks scientific exactness. The Longman suggests a somewhat better definition: 'the ability of something to stretch and go back to its usual length or size'. The

TSF contains an even more exact formulation: 'the ability of a material or product to undergo more or less significant springy reversible deformations without destruction' ['спроможність матеріалу або виробу зазнавати більш або менш значних пружних зворотливих деформацій без руйнування']. This last definition should be used in wordnets.

Flexure, flexion, flection (i110688): 'the state of being flexed (as of a joint)'. This definition refers the concept to the corresponding verb and is, therefore, not sufficient. The TSF defines this concept as 'a type of deformation characterized by a change in the curvature of the axis (bar, balk, rod) or the middle surface (plates, shells) under the influence of external force or temperature' ['вид деформації, що характеризується зміною кривизни осі (бруса, балки, стержня) або серединної поверхні (пластинки, оболонки) під дією зовнішньої сили або температури'] that is much more informative and can be recommended.

Fluorescence (i97800): 'light emitted during absorption of radiation of some other (invisible) wavelength'. This definition does not emphasize essential features of fluorescence and refers to light which is correct but not complete. The TSF gives a definition that refers fluorescence to luminescence and describes its essential features and, therefore, may be recommended: 'short-term luminescence with decay time of 10^{-8} – 10^{-9} sec' ['короткочасна люмінесценція з часом загасання 10^{-8} – 10^{-9} сек'].

Loupe (i55838): 'small magnifying glass (usually set in an eyepiece) used by jewelers and horologists'. This definition is too narrow. The TSF gives a more general and comprehensive definition, which is recommended: 'an optical system consisting of a lens or several lenses, designed to observe small objects located at a finite distance, under an increased angle of view' ['оптична система, що складається з лінзи або кількох лінз, призначена для спостереження дрібних предметів, розташованих на скінченній відстані, під збільшеним кутом зору'].

Resistance, impedance, ohmic resistance, resistivity, electric resistance, electrical resistance (i98048): 'a material's opposition to the flow of electric current; measured in ohms'. This definition makes no distinction between direct and alternating currents, which is crucial here. In fact, impedance is analogical to resistance but represents a different concept. The Longman defines impedance as 'a measure of the power of a piece of electrical equipment to stop the flow of an alternating current'. The TSF gives a more precise definition of impedance: 'an analogue of electrical resistance for harmonic processes' ['аналог електричного опору для гармонічних процесів'], which is recommended.

Selectivity (i64341): 'the property of being selective'. The TSF gives another definition used in optics that should be also present in wordnets: 'the ability of a spectral device to distinguish narrow spectral intervals from a continuous spectrum' ['спроможність спектрального прилада виділяти вузькі спектральні інтервали з суцільного спектру'].

Sensing element, sensor, detector (i52820): 'any device that receives a signal or stimulus (as heat or pressure or light or motion etc.) and responds to it in a distinctive manner'. This formulation is too broad, and the term "detector" does not have this sense. The Longman defines sensor similarly but incompletely, as 'a piece of equipment used for discovering the presence of light, heat, movement etc.', and detector differently, as 'a machine or piece of equipment that finds or measures something'. The TSF defines this concept as 'a converter of a controlled or regulated value into an output signal, convenient for remote transmission and further processing' ['перетворювач контрольованої чи регульованої величини у вихідний сигнал, зручний для дистанційної передачі та подальшої обробки'], and detector as 1 'a device for detecting certain particles or radiations' ['прилад для виявлення тих чи інших частинок або випромінювань'] and 2 'a device for converting modulated high-frequency oscillations into oscillations with a modulation frequency' ['прилад для перетворення модульованих високочастотних коливань у коливання з частотою модуляції']. These definitions are recommended for the use in wordnets.

Set, place, pose, position, put, lay (i29197): 'put into a certain place or abstract location'. This definition is not well-formed because it repeats the defined concept ("put"). The Longman avoids this inconsistency: 'to move something to a particular place or position, especially using your hands'. This definition should be recommended for wordnets.

Undulation, wave (i75294): '(physics) a movement up and down or back and forth'. This definition is too simplified to convey notable information. The Longman proposes a more relevant but too narrow definition: 'a line of raised water that moves across the surface of the sea'. The more strict definition is available in the "Explanatory dictionary of physics": 'a change in some set of physical quantities (fields), capable of moving (spreading), moving away from the place of its origin, or oscillating within limited regions of space' ['зміна деякої сукупності фізичних величин (полів), здатна переміщуватися (поширюватися), віддаляючись від місця її виникнення, або коливатися всередині обмежених областей простору'] which we recommend.

Incomplete OEWN definitions

Flint (i115216): 'a hard kind of stone; a form of silica more opaque than chalcedony'. The Longman distinguishes two meanings: 1 'a type of smooth hard stone that makes a small flame when you hit it with steel'; 2 'a piece of this stone or a small piece of metal that makes a small flame when you hit it with steel' and the TSF adds a sense 'optical glass that contains lead oxide and thus has a refractive index and its dispersion greater than, for example, crown' ['оптичне скло, яке містить окис свинцю і завдяки цьому має показник залому та його дисперсію більші, ніж, наприклад, крон']. The wordnet definitions should include all these meanings.

Graphite, black lead, plumbago (i114791): 'used as a lubricant and as a moderator in nuclear reactors'. This is not a complete definition, but rather a part of it. Another and essential part is available in the "Longman Dictionary of Contemporary English": 'a soft black substance that is a kind of carbon, used in pencils, paints, and electrical equipment'. The combined formulation for graphite is recommended: 'a soft black substance that is a kind of carbon, used in pencils, paints, and electrical equipment, and that serves as a lubricant and as a moderator in nuclear reactors'.

Median (i68434): 'the number midway between the two middle numbers in a series containing an even or odd number of items'. This definition is too complicated and encompasses only one sense of this polysemant. In comparison, the Longman distinguishes three senses of this noun: 1 'a narrow area of land that separates the two sides of a big road in order to keep traffic travelling in different directions apart'; 2 'the middle number of measurement in a set of numbers or measurements that have been arranged in order'; 3 'a line passing from one of the points of a triangle to the centre of the opposite side'. The TSF discerns also three senses of this term: 1 'the line that bisects the opposite side of the polygon' ['лінія, що ділить протилежну сторону багатокутника навпіл'], which is absent in OEWN and is more general than sense 3 of Longman; 2 'a point with the smallest sum of distances from given points' ['точка з найменшою сумою відстаней від заданих точок'], which is absent in OEWN and Longman; 3 'the value in the middle of a certain set of values' ['середнє значення певного набору величин'], which correlates with the OEWN and Longman definitions but is more comprehensive and clear. It is recommended to use the definitions of median given in the "Explanatory dictionary of physics".

Oscillator (i56791): 'generator that produces sonic oscillations or alternating current'. This treatment is too narrow. The Longman contains a more exact formulation: 'a machine that produces electrical oscillations', whereas the TSF adds two more definitions describing two other meanings of this term: 1 'a system (or a material point) carrying out oscillating periodic motion near the position of stable equilibrium' ['система (або матеріальна точка), що здійснює коливальний періодичний рух біля положення стійкої рівноваги']; 2 'exciter (source) of oscillations' ['збуджувач (джерело) коливань']. The last three definitions should be recognized also in wordnets.

Photometry (i40656): 'measurement of the properties of light (especially luminous intensity)'. There is also another meaning presented in the "Explanatory dictionary of physics": 'a section of physical optics dedicated to the measurement of electromagnetic radiation in the optical range' ['розділ фізичної оптики, присвячений вимірюванню електромагнітного випромінювання оптичного діапазону'] — that should be included in wordnets.

Radiometer (i57900): 'meter to detect and measure radiant energy (electromagnetic or acoustic)'. The TSF three more definitions that should be present in

wordnets: 1 'a device for measuring the energy of electromagnetic radiation, based on its thermal action' ['прилад для вимірювання енергії електромагнітного випромінювання, заснований на його тепловій дії']; 2 'a radio telescope receiver' ['приймальний пристрій радіотелескопа']; 3 'a device for measuring the activity of radioactive sources' ['прилад для вимірювання активності радіоактивних джерел'].

Spectrometry, spectroscopy (i38839): 'the use of spectrometers to analyze spectra'. This definition is incomplete and makes no difference between different notions of spectrometry and spectroscopy. The TSF defines spectrometry as 'a set of methods and theory of measuring the spectra of electromagnetic radiation and studying the spectral properties of substances and solids in the optical range of wavelengths' ['сукупність методів і теорія вимірювання спектрів електромагнітного випромінювання та вивчення спектральних властивостей речовин і тіл в оптичному діапазоні довжин хвиль'] and proposes two senses for spectroscopy: 1 'a field of physics devoted to the study of the distribution of the intensity of electromagnetic radiation by wavelengths or frequencies' ['область фізики, присвячена дослідженню розподілу інтенсивності електромагнітного випромінювання за довжинами хвиль або частотами']; 2 'study of decompositions into the spectrum' ['дослідження розкладів у спектр']. These two different notions should be treated in wordnets accordingly.

Transparency, transparence (i98144): 'permitting the free passage of electromagnetic radiation'. The Longman does not have an entry "transparence" and proposes three senses for "transparency": 1 'a sheet of plastic or a piece of photographic film through which light can be shone to show a picture on a large screen'; 2 'the quality of glass, plastic etc. that makes it possible for you to see through it'; 3 'the quality of being easy to understand or know about'. Here, the second sense gives qualitative description of this concept. In turn, the TSF presents more strict definition: 'a value that shows how much of the radiation flux (or for visible light – the luminous flux) falling on a certain surface passes through a layer of unit thickness without changing its direction' ['величина, що показує, яка частина потоку випромінювання (або для видимого світла – світлового потоку), що падає на деяку поверхню, проходить без зміни напрямку через шар одиничної товщини'], which is recommended.

The proposed definition amendments will, accordingly, cause the changes in hypo-hypernymy relations that should be correctly displayed in wordnets.

The relative number of such inconsistencies is small because the total number of analyzed synsets in OEWN is 3 220. However, each third definition therein requires minor refinement.

The described procedure may be executed both automatically — through hypothetical corresponding tool based on the findings of (Vakulenko 2019; 2021) — or manually. Human validation by an experienced lexicographer is required.

5. Discussion

The analyzed problems are typical for wordnets created on the basis of the "merge" and "combined" approaches. The "expansion" approach alone does not have these inconsistencies because the resulting data in this case come from only one source — pIWordNet (Rudnicka et al. 2019), which is linked to OEWN. However, its disadvantage is that in most cases, the resulting synsets and their semantic structures do not correspond to those inherent to the target language, not to mention individual lexical units, whose machine translation is often incorrect today, especially with regard to the Ukrainian language. This results in the need for tedious post-processing that cannot yet be performed by a machine.

The profound reason for these inconsistencies between concepts and their verbal designations lies in asymmetrical dualism of the linguistic sign that is the fundamental property of language. Asymmetrical dualism of the linguistic sign means that the word meaning (signifié) and its written form (signifiant) are set in an unstable balance and display dynamic behavior. The first is inclined to be expressed by other word forms, thus giving rise to synonymy, while the second tends to acquire additional meanings (see Karcevskij 1929). This continuous emergence of additional meanings cannot be properly tracked by conventional lexicography methods.

Fellbaum (2006: 665) makes a reserving remark about the interchangeability of synset members in different contexts, indicating that there exist synset members that cannot be substituted in all contexts. However, if synonymy is considered as a phenomenon formed at a higher, contextual level, it may be concluded that lexical units that cannot be substituted in all contexts, should belong to different synsets. From this standpoint, a revised definition of synonyms strictly requiring their interchangeability in all contexts should be adhered to.

Regarding different definitions of the same lexeme, a reasonable question may be asked: how crucial is the given form of a definition for a semantic load of the defined unit? In other words, how do the differences in definitions available in different sources affect the sense of the lexeme in question?

A metric method to measure semantic distances between lexemes was applied to build semantic fields using a recursive procedure that represents a lexeme's meaning by its semes taken from a dictionary definition (Vakulenko 2021). It was observed that two types of lexicographic effects appear in the Merriam-Webster dictionary⁹: the horizontal crossing of different semantic trees when some semes belonging to different trees, are the same at certain semantic depth; and the vertical looping (appearance of lexicographic hypercycles) when the same seme appears at some depth of the given semantic tree. A similar lexicographic effect was reported earlier in Shyrov (2018: 12-18) on the material of the "Dictionary of Ukrainian Language in 20 Volumes" where the following semantic hypercycle was found: *prylad* 'instrument' – *prystrij* 'device' (hypernym) – *obladnannja* 'equipment' (hypernym) – *prylad* 'instrument' (meronym). Such hypercycles appear in sufficiently large dictionaries and manifest the rise

of semantic structures that are conditioned by vertical and horizontal semantic correlations between concepts. The existence of these structures indicates that the resulting semantic fields display a certain level of stability against different formulations in definitions.

The outcomes of the recent research on concept ontology (Bense 2021) suggest that complex concepts can finally be decomposed into a significantly smaller number of elementary concepts displaying more semantic clarity, which are the same across analyzed languages. This fact also supports the assumption that the semantic load of a given lexeme is less dependent on the exact formulation of the corresponding concept.

6. Prospects of future research

Some related questions that should be examined in more detail in future investigations are briefly discussed here.

1. The notion of "meaning dimensions" (Löbner 2013) has a major sense in semantics because it sets the weight of any individual meaning. The further development if this idea leads to the conclusion that any sense of a linguistic unit can be presented in the form of its semantic field represented by a sum of elementary meanings taken with corresponding weight coefficients (see Vakulenko 2021). In other words, a proper account of "meaning dimensions" requires basing on their weight coefficients which determine the importance of these dimensions.

The definition of synonyms by Löbner (op. cit.) uses a vague concept of "sameness" and tempts a logical question of what senses are considered the "same", or to what extent of precision the lexical units must have the "same" meaning, and what the tools to measure this "sameness" could be. This fuzzy definition makes the synsets themselves ambiguous and unclear. On the other hand, a mathematically strict approach to calculate semantic distances between lexical units based on their definitions (Vakulenko 2019) may resolve this issue. The possible use of underlying mathematics to determine and measure synonymy would be promising for further development of wordnets.

2. The semantic equivalence of the compound synonyms *electron lone pair* and *lone-pair electrons* (TSF: 394) is due to syntactic reasons. This kind of similarity may be displayed also by corresponding contextualized embeddings as described in (Liu et al. 2020; Lee 2021; Loureiro et al. 2021). However, the method described in Section 3 to refine relevant semantics seems to be much more straightforward and clearer.
3. The development of semantic structures obtained within the "merge" approach, requires assignment of semantic relations not displayed in the lexicographic data sources. For example, "A Multilingual Dictionary on Phonetics and Phonology" (Yurtbaşı 2017a; 2017b; Vakulenko 2018: 49-55) and "A Concise

Dictionary of Synonyms of the Ukrainian Language"¹⁰ contain synsets but have no information about hyper-/hyponymy, mero-/holonymy, antonymy, etc. In this case, tools for automatically assigning LSRs are useful, such as the tool for building hypernymy described in Romanyshyn et al. 2024.

However, such a semantic hierarchy is disrupted in multiword expressions where the overall meaning cannot be inferred from the meanings of individual components (Voula 2023).

4. AI technology has been successfully applied to the task of writing dictionary definitions (De Schryver 2023: 381) and text generation, including Ukrainian (*Ukrainian Text Generator*). This makes it possible to expect that this technology will be able to solve also the tasks described above.

7. Conclusion

This study has demonstrated that the development of the Ukrainian Wordnet (Ukrajinet) under the "merge" and "combined" approaches inevitably exposes inconsistencies between lexicographic sources and wordnet structures, especially regarding synonymy and hyper-/hyponymy. What is new and significant here is the identification of systematic mismatches between traditional dictionary definitions and their Open English Wordnet counterparts, and the formulation of a reproducible procedure for their correction. By applying methods of terminology science and comparative lexicography, this work clarified how missing or imprecise definitions distort conceptual hierarchies and proposed a structured post-processing framework that restores their integrity.

These findings contribute to the broader field of computational lexicography by highlighting that semantic refinement is not merely an editorial task but an essential step toward ensuring conceptual accuracy and interoperability of multilingual lexical databases. The recognition that synonymy and hyper-/hyponymy relations often depend on subtle shifts in definitions provides a pathway for developing more semantically consistent ontologies. In practical terms, the proposed procedures can inform the semi-automatic refinement of lexical databases, combining algorithmic processing with expert validation.

However, several limitations remain. The current refinement was performed on a subset of data and focused primarily on the domains of physics and phonetics. Extending this methodology to other domains will require additional testing and refinement of the semantic alignment rules. Moreover, while partial automation is achievable, full automation of lexicographic post-processing still depends on advances in AI-based semantic modeling.

The practical implications of this research lie in improving the quality of multilingual wordnets, supporting tasks such as machine translation, semantic search, and ontology construction. Future research should focus on implementing computational tools capable of identifying definitional inconsistencies automatically, experimenting with weighted semantic fields, and integrating contextual embeddings to enhance synonymy detection. Further theoretical work

on semantic metrics and concept ontologies may also provide mathematical foundations for refining lexical relations.

Ultimately, this study underscores that the future of lexicography lies in the synergy of human expertise and computational precision. Wordnets are not static repositories but evolving models of linguistic and conceptual knowledge. The refinement procedures outlined here contribute to moving closer to a truly interoperable and semantically transparent lexicon — a cornerstone for intelligent language technologies of the future.

8. Acknowledgments

The author is grateful to Prof. Melanie Siegel for useful comments and suggestions that helped to improve the content of this article.

9. Endnotes

1. <https://en-word.net/>
2. <https://ukrajinet.yoda.media.h-da.de/>
<https://github.com/hdaSprachtechnologie/ukrajinet>
https://code.fbi.h-da.de/ukrajinet/ukrajinet-dev/-/blob/main/20230728_merged_ukrajinet.xml
3. Vakulenko, M.O and O.V. Vakulenko. 2008. *Tlumachnyj slovnyk iz fizyky* [Explanatory Dictionary on Physics]. Kyjiv: VPC "Kyjivsjkyj universytet". 767 p.
4. Vakulenko, M.O. and O.V. Vakulenko. 2017. *Slovnyk fizychnykh terminiv-synonimiv* [Dictionary of Physical Synonymous Terms]. Kyjiv: VPC "Kyjivsjkyj universytet". 191 p.
5. Rusanivsjkyj, Vitalij (Ed.). 2010-. *Slovnyk ukrajinsjkoji movy v 20 tomakh* [Dictionary of Ukrainian Language in 20 Volumes]. Kyjiv: Naukova dumka.
6. *Longman Dictionary of Contemporary English*. 2003. [Della Summers (ED.).] Harlow, Essex: Pearson Education.
7. Sube, Ralf and Günther Eisenreich. 1973. *Wörterbuch Physik. A–M. Englisch–Deutsch–Französisch–Russisch*. Zürich/Frankfurt am Main: Verlag Harri Deutsch.
8. Greulich, Walter and Dirk Meenenga. 1997. *Wörterbuch der Physik: Englisch–Deutsch/Deutsch–Englisch*. Heidelberg/Berlin: Spektrum, Akademischer Verlag.
9. *Dictionary by Merriam-Webster*. 2023.
<https://www.merriam-webster.com/>
10. Derkach, Pylyp. 1960. *Korotkyj slovnyk synonimiv ukrajinsjkoji movy* [A Concise Dictionary of Synonyms of the Ukrainian Language]. Kyjiv: Radjansjka shkola.
https://chtyvo.org.ua/authors/Derkach_Pylyp_Maksymovych/Korotkyi_slovnyk_synonimiv_ukrajinsjkoji_movy/

References

- Bense, Hermann.** 2021. A Formal Ontology of Concepts for Deep Semantic Searching. *KONT-21 Knowledge — Ontology — Theory, 8th International Conference, 8–12 November 2021, Novosibirsk*.
https://www.ontology4.us/download/dot/ontopage/sub_Bens2021d.pdf

- Bokshanj, Ghalyna.** 2021. Synonimija v terminosystemi anghlomovnykh naukovykh tekstiv ekonomichnogho sprjamuvannja [Synonymy in Terminology System of English-language Scientific Texts in the Field of Economics]. *Pivdennyj arkhiv (filologichni nauky)* 86: 54-57 (in Ukrainian). DOI: [10.32999/ksu2663-2691/2021-86-8](https://doi.org/10.32999/ksu2663-2691/2021-86-8)
- Bond, Francis, Piek Vossen, John P. McCrae and Christiane Fellbaum.** 2016. CILL: the Collaborative Interlingual Index. Fellbaum, Christiane, Piek Vossen, Verginica Barbu Mititelu and Corina Forascu. 2016. *Proceedings of the 8th Global WordNet Conference (GWC)*, Bucharest, Romania, 27–30 January 2016: 50-57. Global Wordnet Association. <https://aclanthology.org/2016.gwc-1.9>
- De Schryver, Giles-Maurice.** 2023. Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography* 36(4): 355-387. DOI: <https://doi.org/10.1093/ijl/ecad021>
- Fellbaum, C.** 1998. *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.
- Fellbaum, C.** 2006. WordNet(s). Keith Brown (Editor-in-Chief). 2006. *Encyclopedia of Language & Linguistics*. Second Edition. Volume 13: 665-670. Oxford: Elsevier.
- Halas Popović, Ana.** 2023. Synonymy from a Prototype Theory Perspective and its Symbiosis with Polysemy: Towards a New Dictionary of Synonyms. *Lexikos* 33: 1-23. DOI: <https://doi.org/10.5788/33-1-1773>
- Karcevskij, Sergej.** 1929. Du dualisme asymétrique du sign linguistique [On the Asymmetric Dualism of the Linguistic Sign]. *Travaux du Cercle linguistique de Prague* 1: 88-92. <https://www.degruyter.com/document/doi/10.7560/780439-004/html>
- Lee, Younghoon.** 2021. Systematic Homonym Detection and Replacement Based on Contextual Word Embedding. *Neural Processing Letters* 53: 17-36.
- Liu, Qi, Matt J. Kushner and Phil Blunsom.** 2020. A Survey on Contextual Embeddings. ArXiv.2003.07278v2 [cs.CL]. DOI: <https://doi.org/10.48550zrXiv.2003.07278>
- Loureiro, Daniel, Kiamehr Rezaee, Mohammad Taher Pilehvar and Jose Camacho-Collados.** 2021. Analysis and Evaluation of Language Models for Word Sense Disambiguation. *Computational Linguistics* 47(2): 387-443. DOI: https://doi.org/10.1162/coli_a_00405
- Löbner, Sebastian.** 2013. *Understanding Semantics*. Second Edition. New York: Routledge.
- Maziarz, M., M. Piasecki and S. Szpakowicz.** 2013. The Chicken-and-Egg Problem in Wordnet Design: Synonymy, Synsets and Constitutive Relations. *Language Resources and Evaluation* 47(3): 769-796.
- McCrae, John Philip, Alexandre Rademaker, Ewa Rudnicka and Francis Bond.** 2020. English Wordnet 2020: Improving and Extending a Wordnet for English Using an Open-source Methodology. Declerk, Thierry, Itziar Gonzalez-Dios, German Rigau. 2020. *Proceedings of the LREC 2020 Workshop on Multimodal WordNets (MMW2020)*, Marseille, France, May 2020: 14-19. The European Language Resources Association (ELRA). <https://aclanthology.org/2020.mmw-1.3/>
- Romanyshyn, Nataliia, Dmytro Chaplynskyi and Mariana Romanyshyn.** 2024. Automated Extraction of Hypo-Hypernym Relations for the Ukrainian WordNet. *The Third Ukrainian Natural Language Processing Workshop (UNLP) @LREC-COLING-2024, 25 May 2024*: 51-60. ELRA Language Resource Association.

- Rudnicka, Ewa, Maciej Piasecki, Francis Bond, Łukasz Grabowski and Tadeusz Piotrowski.** 2019. Sense Equivalence in plWordNet to Princeton WordNet Mapping. *International Journal of Lexicography* 32(3): 296-325.
DOI: <https://doi.org/10.1093/ijl/ecz004>
- Shyrov, V.A. (Ed.).** 2018. *Linghivistychno-informacijni studiji. Praci Ukrajinsjkogho movno-informacijnogho fondu NAN Ukrainy: u 5 tomakh, tom 3. Tlumachna leksykoghracija, knygha 2. Systemna semantyka tlumachnykh slovnykiv* [Linguistic Information Studies. Works of Ukrainian Lingua-Information Fund of National Academy of Sciences of Ukraine: in 5 Volumes, Vol. 3. Explanatory Lexicography, Book 2. Systemic Semantics of Explanatory Dictionaries] (in Ukrainian). Kyjiv: Ukrajinsjkyj movno-informacijnyj fond.
- Siegel, Melanie and Francis Bond.** 2021. OdeNet: Compiling a German Wordnet from Other Resources. Vossen, Piek and Christiane Fellbaum (Eds.). 2021. *Proceedings of the 11th Global Wordnet Conference (GWC 2021), Potchefstroom, South Africa, 18–21 January 2021*: 192-198.
<https://www.aclweb.org/anthology/2021.gwc-1.22>
- Ukrainian Text Generator.** Linguanex.
<https://lingvanex.com/services/ukrainian-text-generator/>
- Vakulenko, Maksym.** 2018. *From Terminology-Vocabulary to Terminology-Science: A Ukrainian Trend*. Beau Bassin: LAP.
- Vakulenko, Maksym.** 2019. Calculation of Semantic Distances between Words: From Synonymy to Antonymy. *Journal of Quantitative Linguistics* 26(2): 116-128. (Published online: 28 May 2018.)
<https://doi.org/10.1080/09296174.2018.1452524>
- Vakulenko, Maksym.** 2021. From Semantic Metrics to Semantic Fields. *Proceedings of the 2021 IEEE 16th International Conference on Computer Science and Information Technologies (CSIT), 22–25 September 2021, Lviv, Ukraine*: 44-47.
DOI: 10.1109/CSIT52700.2021.9648675
- Voula, Giouli.** 2023. A Model for Representing the Semantics of MWEs: From Lexical Semantics to the Semantic Annotation of Complex Predicates. *Frontiers in Artificial Intelligence* 6.
DOI: <https://doi.org/10.3389/frai.2023.802218>
- Yurtbaşı, Metin.** 2017a. The French Phonological Corpus with Equivalents in English, German and Turkish, as Part of a New Multilingual Dictionary Project. *International Journal of Arts, Humanities, and Social Sciences* 2(11): 27-77.
<http://www.ijahss.com/vol2-issue11.html>
- Yurtbaşı, Metin.** 2017b. A Multilingual Dictionary also Containing Qazaq Terminology in Phonetics and Phonology. *International Journal of Arts, Humanities, and Social Sciences* 2(11): 78-103.
<http://www.ijahss.com/vol2-issue11.html>

Out-of-the-Box GPTs for Lexicography

Gilles-Maurice de Schryver, *BantUGent & LT³, Ghent University,
Ghent, Belgium; and Department of African Languages,
University of Pretoria, Pretoria, South Africa*
(gillesmaurice.deschryver@UGent.be)
(<https://orcid.org/0000-0001-7272-9878>)

Abstract: A detailed analysis is offered of how ChatGPT-3.5 has treated the world's languages for lexicographic purposes, starting with hyper-visible English, moving to semi-visible Portuguese, and ending with low-visible Bantu languages. Case studies for the latter group are region-visible Swahili, nation-visible Xhosa and local-visible Lusoga. This impact-based scale is reflected in the uptake of GenAI technology for lexicography. The results of using out-of-the-box ChatGPT-3.5 are fantastic for English, fake for Portuguese, and simply fail for the Bantu languages. But there is a way out, and the seeds for that are sown in this article.

Keywords: GENERATIVE ARTIFICIAL INTELLIGENCE (GENAI), LARGE LANGUAGE MODELS (LLMs), GENERATIVE PRE-TRAINED TRANSFORMER (GPT), OUT-OF-THE-BOX GPTs, LEXICOGRAPHY, DICTIONARIES, LEXICOGRAPHERS, ENGLISH, PORTUGUESE, BANTU, SWAHILI, XHOSA, LUSOGA

Obufunze: GPT ey'omutindo eya GenAI mw'isomo ly'okuwandiika amawanika. Okutangaazibwa ku ngeli ChatGPT-3.5 bw'ekozeseibwa mu nnimi dh'ensi yoonayoona mu kugezaaku okutuukiliza ebigendelelwa by'okuwandiika amawanika, kuweebwa mu lupapula luno. Okutangaaza kuno kutandiikila ku nkozesa eyolekebwa ku lulimi oluganzi mu nsi yoonayoona olw'Olungeleza, kw'amala kwaizibwa ku nnimi engazimuuku akatono ng'Olupootigo, kw'asebyayo enkozesa eyeeyolekela mu nnimi edhikozesebwa abantu abatono einho ng'edho edhigwa mu luse lwa Bantu. Ennimi dha Bantu edhiliku enkozesa entono einho edhigelaagela-nhizibwa ku y'edho emba oba enjanziku n'omuli, olulimi lw'Oluswayili, Oluxhosa n'Olusoga. Ennimi dhino dhonsatule n'edhisinziilwaku okulaga tekinologiya wa GenAI bwe yeeyolekelamu ng'akozesebwa mu milimu gy'okuwandiika amawanika. Kizuuse nti enkozesa ya tekinologiya eno nambulukufu inho mu milimu egikola ku mawanika g'Olungeleza, aye, ya kibogwe ku nnimi enjanziku akatono ng'Olupootigo. Bwe kituuka ku nnimi dh'omu luse lwa Bantu kabili enkola eno eililila inho y'aba oti kumpi tekelela ilala. Aye, wakaaliwooku eisuubi eliva mu byandhulwa mu kitundu ky'olupapula luno eky'okubili. Mu lupapula olw'okubili n'omuli ebyava mu kwongela okunoonheleza ku ngeli eyinza okukozesebwa okuziba olutanda oluli wagati w'ennimi dh'ensi yoonayoona enjanzu, n'edhitali, aye nga dhoonadhoona dhilubilila okukoseza GenAI okuwandiika amawanika.

Ebigambo ebikulu: KALIMAGEZI-KISIBUKILA, OKUTETENKELEZA KW'ENNIMI ENNHINGI, KISIBUKILA AKAALI KUKENGULWA, GPT EY'OMUTINDO, EISOMO LY'OKUWA-

NDIIKA AMAWANIKA, AMAWANIKA, ABAWANDIISI B'AMAWANIKA, OLUNGELEZA, OLUPOOTIGO, BANTU, OLUSWAYILI, OLUXHOSA, OLUSOGA

1. Prologue

In De Schryver (forthcoming-a, forthcoming-b) a sample of 100 contributions on Generative Artificial Intelligence (GenAI) in lexicography was studied, covering the period from the release of ChatGPT on 30 November 2022 up to 31 May 2025. Over those two years and a half, all but one of the contributions simply used *out-of-the-box* Generative Pre-trained Transformer (GPT) models. Those types of Large Language Models (LLMs) are known for their ability to generate sophisticated human-like texts. Just a single one of the 100 contributions moved to the next level for lexicography, namely to a *customised* GPT model (De Schryver 2024a). However, and despite the expressed wishes about it at gatherings on digital lexicography, starting with eLex 2023, not a single attempt has been made to date to move to the next-next level, namely to a *fine-tuned* GPT model. In the present article the first step is revisited for ChatGPT, and in a next article (De Schryver 2025) the second step is revisited with attention to both OpenAI's ChatGPT and Google's Gemini, expanding the four-page abstract on it prepared (but not delivered) for a workshop on LLMs and lexicography at Euralex 2024 (De Schryver 2024c).

2. Out-of-the-box GPTs

2.1 Hyper-visible English

By and large, the most successful experiments with LLMs for lexicography have been reported on for applications using English (1.45 billion speakers; spoken in 59 sovereign countries and 27 non-sovereign entities). See especially De Schryver (2023b), Lew (2023), Rees and Lew (2024), Lew (2024), Ptasznik et al. (2024) and Lew et al. (2024). At the same time, some colleagues correctly pointed out a few caveats, see especially McKean and Fitzgerald (2023), Jakubíček and Rundell (2023) and Nichols (2023). That said, virtually all scholars agree that it is not a good idea not to bring in today's GenAI for English lexicography.

2.2 Semi-visible Portuguese

The results of the first in-depth and detailed attempt to study the lexicographic potential of using ChatGPT for a language other than English, were presented at the inaugural Americalex-S conference in São Paulo, Brazil (De Schryver 2023a). The assumption was that ChatGPT would do an excellent job for the rather large language Portuguese (260 million speakers; spoken in 10 coun-

tries and territories). Following the success of using ChatGPT-3.5 for compiling a monolingual English COBUILD-style learners' dictionary — demonstrating its incredible performance on lemmas from a variety of word classes: verbs, nouns, conjunctions, adjectives and interjections, as well as metaphorical meanings and idiomatic expressions (see De Schryver 2023b: supplementary material) — the first goal was to test ChatGPT-3.5 on a monolingual Portuguese COBUILD-style learners' dictionary. There was an ongoing student strike on campus during the conference, so the polysemous lemma *strike* was chosen, and put into a Portuguese translation of the same prompt which had been used for English in De Schryver (2023b: supplementary material), as shown in Figure 1. Here and elsewhere, English translations of the Portuguese are to be found in Addendum A. Note that in De Schryver (2023b) the goal was to arrive at a single prompt with which each type of lemma, no matter its part of speech or any other aspects, may be generated. The 'prompt engineering' thus preceded the prompt eventually proposed, with which an entire COBUILD dictionary could be compiled. Copies of De Schryver's (2023b) prompt have subsequently been used by colleagues, such as Zhong (2025) using DeepSeek, and also in translated form, such as in Spanish by Fuertes-Olivera (2024), or in Lithuanian by Pavydytė (2024).

Default (GPT-3.5)



Por favor, forneça um verbete de dicionário no estilo COBUILD para "strike". Comece com as informações de frequência na parte superior. Cada sentido deve estar em um bloco numerado. Cada bloco começa com a classe gramatical e as formas morfológicas do respectivo sentido. Em seguida, há uma definição de sentido e exemplos de sentido que ilustram o uso e o significado de cada sentido específico. Para as frases de exemplo, certifique-se de usar diferentes estruturas de frases, referindo-se a diferentes pessoas; refira-se a situações passadas, presentes e futuras; varie frases de exemplo longas e curtas; e inclua outras elaborações para que você, caro ChatGPT, produza material de dicionário de alta qualidade para mim.

Figure 1: ChatGPT-3.5 prompt to obtain a COBUILD-style entry for a monolingual Portuguese learners' dictionary, here for the lemma *strike*

Even though the prompt was written in Portuguese, the reply from ChatGPT was entirely in English, as seen in Figure 2 (only showing the first part, the start of the verbal use).

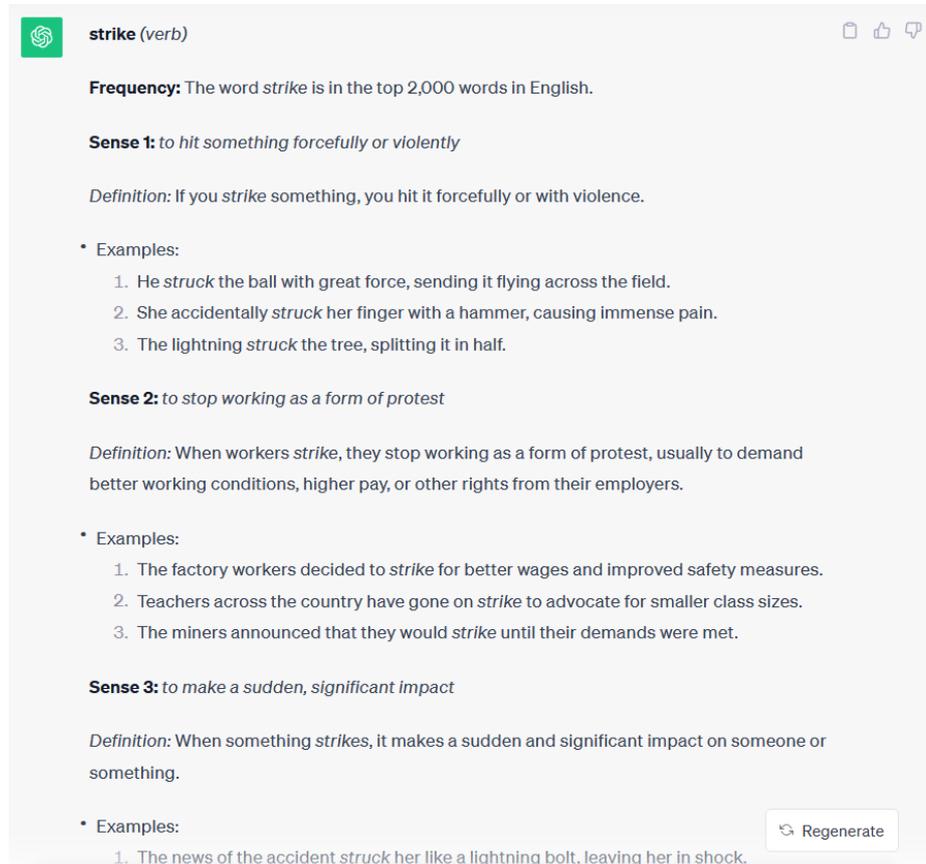


Figure 2: Start of the ChatGPT-3.5 output for the prompt seen in Figure 1

This was followed up with: 'Thanks, but actually I wanted to see a monolingual dictionary entry for the Portuguese entry '*greve*' (which is 'strike' in English)', after which ChatGPT replied with what is seen in Figure 3.

This time, ChatGPT presented a noun (correct), with an indication that it is feminine (correct), in a wider frequency band compared to English (ambiguous), a good definition (but not a COBUILD-style one), acceptable examples and an interesting encyclopaedic note. Observe that examples 1 and 2 in Figure 3, are very similar to examples 1 and 2 under the verbal sense 2 seen in English in Figure 2, except that the second one there is not even a verb, but part of a verb phrase where *strike* is used as a noun (so an error). From the very first reply by ChatGPT onwards, and now reinforced by the similar examples in the second reply, one starts wondering if one is not dealing with translation issues (in the sense that the model generates everything in and for English first, and only then maps the English onto another language as if it were English).



Obrigado, mas na verdade eu queria ver um verbete de dicionário monolíngue para o verbete português "greve" (que é "strike" em inglês).

Peço desculpas pela confusão anterior. Aqui está um verbete de dicionário monolíngue para a palavra "greve" em português:

greve (*substantivo feminino*)

Frequência: A palavra *greve* está entre as 10.000 palavras mais frequentes em português.

Definição: A *greve* é uma paralisação organizada por trabalhadores, na qual eles deixam de trabalhar como forma de protesto, geralmente para reivindicar melhores condições de trabalho, salários mais altos, ou outros direitos dos empregadores.

Exemplos:

1. Os funcionários da fábrica entraram em *greve* para exigir um aumento salarial.
2. Os professores decidiram iniciar uma *greve* devido à falta de investimentos na educação.
3. A *greve* dos caminhoneiros causou escassez de suprimentos em todo o país.

Nota: A palavra "greve" é frequentemente usada no contexto de protestos trabalhistas e é um direito garantido por leis trabalhistas em muitos países. Ela pode ser realizada por diversos setores profissionais, incluindo professores, médicos, motoristas, e outros, como um meio de pressionar empregadores ou governos a atenderem às demandas dos trabalhadores.

Regenerate

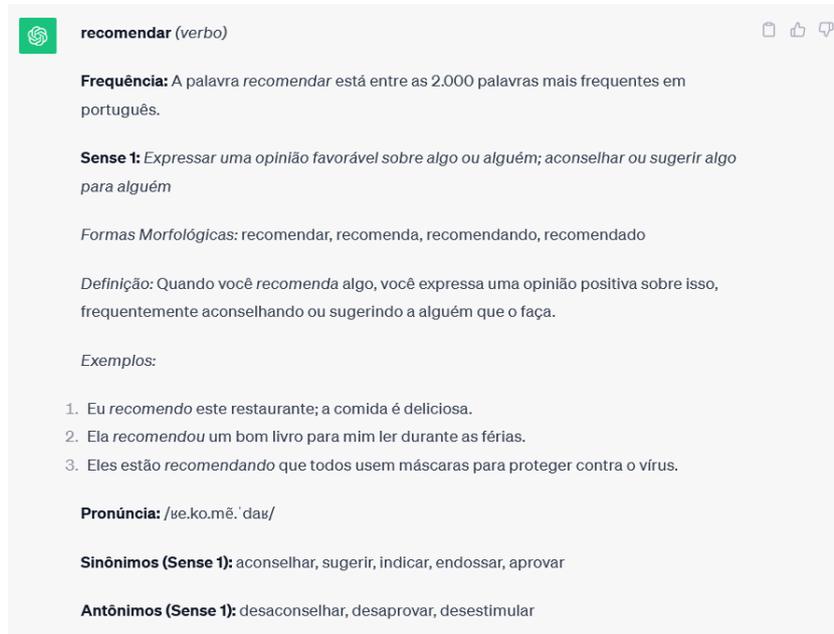
Figure 3: ChatGPT-3.5 output for the noun *greve* 'strike', for a monolingual Portuguese learners' dictionary

Next, the verb *recomendar* 'recommend', being one of the 15 verbs of communication used by Lew (2023), was submitted. The prompt (with an extra line) is shown in Figure 4. The reply by ChatGPT looked rather satisfactory, as may be seen from Figures 5a, b and c.



Por favor, forneça um verbete de dicionário monolíngue no estilo COBUILD para "recomendar". Comece com as informações de frequência na parte superior. Cada sentido deve estar em um bloco numerado. Cada bloco começa com a classe gramatical e as formas morfológicas do respectivo sentido. Em seguida, há uma definição de sentido e exemplos de sentido que ilustram o uso e o significado de cada sentido específico. Para as frases de exemplo, certifique-se de usar diferentes estruturas de frases, referindo-se a diferentes pessoas; refira-se a situações passadas, presentes e futuras; varie frases de exemplo longas e curtas; e inclua outras elaborações para que você, caro ChatGPT, produza material de dicionário de alta qualidade para mim. Inclua também uma pronúncia do verbete usando o IPA e inclua uma lista de sinônimos e antônimos nos sentidos em que essas listas forem relevantes.

Figure 4: ChatGPT-3.5 prompt to obtain a COBUILD-style entry for a monolingual Portuguese learners' dictionary, here for the verb of communication *recomendar* 'recommend'



 **recomendar** (*verbo*) 📄 👍 🗨

Frequência: A palavra *recomendar* está entre as 2.000 palavras mais frequentes em português.

Sense 1: *Expressar uma opinião favorável sobre algo ou alguém; aconselhar ou sugerir algo para alguém*

Formas Morfológicas: recomendar, recomenda, recomendando, recomendado

Definição: Quando você *recomenda* algo, você expressa uma opinião positiva sobre isso, frequentemente aconselhando ou sugerindo a alguém que o faça.

Exemplos:

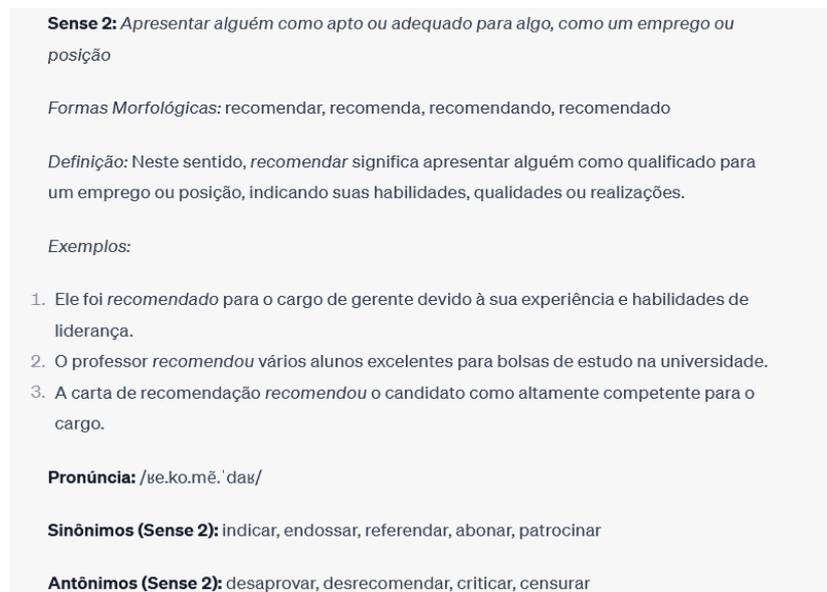
1. Eu *recomendo* este restaurante; a comida é deliciosa.
2. Ela *recomendou* um bom livro para mim ler durante as férias.
3. Eles estão *recomendendo* que todos usem máscaras para proteger contra o vírus.

Pronúncia: /ʁe.ko.mẽ.'daʁ/

Sinônimos (Sense 1): aconselhar, sugerir, indicar, endossar, aprovar

Antônimos (Sense 1): desaconselhar, desaprovar, desestimular

Figure 5a: ChatGPT-3.5 output for the Port. verb *recomendar* 'recommend', # 1, for the prompt seen in Figure 4



Sense 2: *Apresentar alguém como apto ou adequado para algo, como um emprego ou posição*

Formas Morfológicas: recomendar, recomenda, recomendando, recomendado

Definição: Neste sentido, *recomendar* significa apresentar alguém como qualificado para um emprego ou posição, indicando suas habilidades, qualidades ou realizações.

Exemplos:

1. Ele foi *recomendado* para o cargo de gerente devido à sua experiência e habilidades de liderança.
2. O professor *recomendou* vários alunos excelentes para bolsas de estudo na universidade.
3. A carta de recomendação *recomendou* o candidato como altamente competente para o cargo.

Pronúncia: /ʁe.ko.mẽ.'daʁ/

Sinônimos (Sense 2): indicar, endossar, referendar, abonar, patrocinar

Antônimos (Sense 2): desaprovar, desrecomendar, criticar, censurar

Figure 5b: ChatGPT-3.5 output for the Port. verb *recomendar* 'recommend', # 2, for the prompt seen in Figure 4

Sense 3: *Prescrever ou sugerir algo como tratamento ou solução*

Formas Morfológicas: recomendar, recomenda, recomendando, recomendado

Definição: Em termos médicos ou profissionais, *recomendar* envolve prescrever ou sugerir um tratamento, curso de ação ou solução para uma condição ou problema específico.

Exemplos:

1. O médico *recomendou* repouso e medicação para tratar a gripe.
2. O especialista *recomendou* uma dieta equilibrada e exercícios para melhorar a saúde do paciente.
3. Ela *recomendou* que ele consultasse um advogado para resolver a questão legal.

Pronúncia: /ʁe.ko.mẽ.'daʁ/

Sinónimos (Sense 3): prescrever, aconselhar, orientar, sugerir, indicar

Antónimos (Sense 3): desaconselhar, desautorizar, desaprovar, desestimular

Figure 5c: ChatGPT-3.5 output for the Port. verb *recomendar* 'recommend', # 3, for the prompt seen in Figure 4

However, and again — so clearly a serious issue — two of the three definitions are not following the COBUILD-style, despite the explicit request to do so in the prompt. An attempt was then made to even add as second sentence to the prompt: *'The most important aspect of the COBUILD-style is that for the definition, you need to use the full-sentence defining style, which consists of two clauses: the first clause illustrates the usage pattern, the second clause is a paraphrase of meaning'*. But that had no effect whatsoever, despite the fact that merely mentioning COBUILD-style the very first time, resulted in a definition very close to it, see Figure 2 for the verb *strike*: *'If you strike something, you hit it forcefully or with violence'*. Not as perfect as what is seen in an actual COBUILD, but close: *'If you strike someone or something, you deliberately hit them'*.¹ Conclusion: ChatGPT-3.5 is unable to compile proper COBUILD-style entries for a monolingual Portuguese learners' dictionary.

What about a bilingual dictionary? The first focus is on an English to Portuguese entry (in that direction), for the lemma *bank*, with the prompt in English as seen in Figure 6.



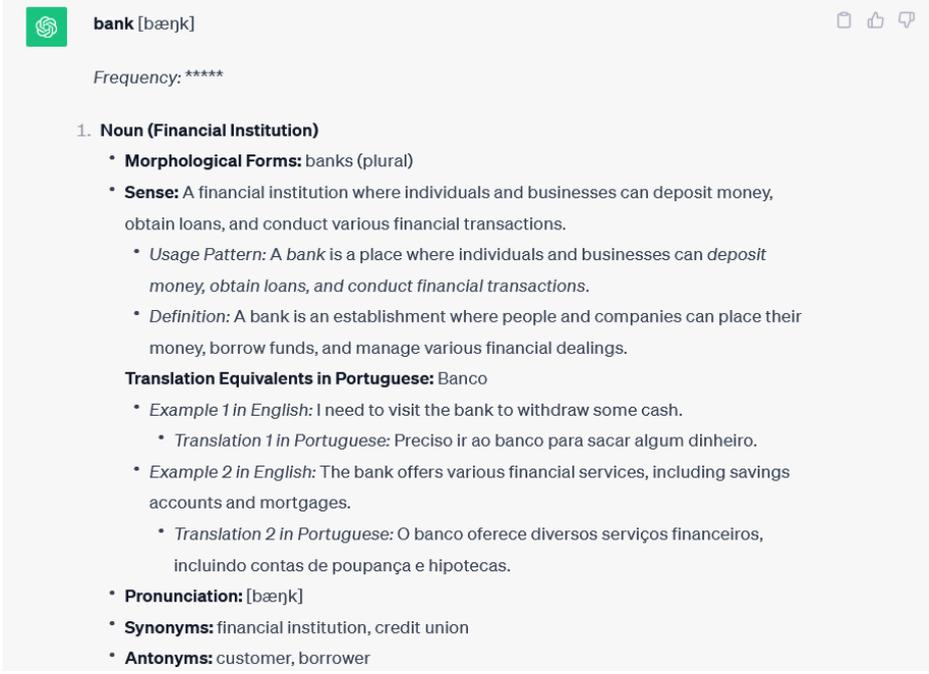
Give me a bilingual English-Portuguese dictionary entry for the word 'bank'. Start with the frequency information at the top. Each sense should be in a numbered block. Each block starts with the part of speech and the morphological forms of the respective sense. This is followed by a sense definition in English, and a translation of that definition into Portuguese. Note that for the definition, you have to use full-sentences, not codes, and these sentences consist of two clauses: the first clause illustrates the usage pattern, the second clause is a paraphrase of meaning. Also list one or more translation equivalents in Portuguese at each sense. Add English examples with their Portuguese translations that illustrate both the use and the meaning of each particular English sense. For the example sentences, make sure to use different sentence structures, referring to different people; refer to past, present, and future situations; vary long and short example sentences; and include other elaborations so that you, dear ChatGPT, produce high-quality dictionary material for me. Also include a pronunciation of the entry using IPA, and include a list of synonyms and antonyms at the senses where such lists are relevant.

Figure 6: Prompt to obtain a COBUILD-style entry for a bilingual English-Portuguese learners' dictionary, here for the lemma *bank*. With also more information on what is meant with a 'COBUILD definition'

The lemma *bank*, a classic since first discussed seven decades ago by Noam Chomsky (1957: 95), and often used by Patrick Hanks ever after (Hanks 1987, 2000, 2013), has multiple nominal and verbal uses. ChatGPT offers three senses, two nominal ones, shown in Figures 7a and b, and one verbal one, shown in Figure 7c.

Clearly, ChatGPT did not understand what is meant with the COBUILD-style full-sentence definition (FSD).² The two clauses became two blocks, one for the usage pattern, the other for the paraphrase of meaning. Despite the fact that three attempts at defining are now made, under the heading 'Sense', the heading 'Usage Pattern' and the heading 'Definition', it is only at the third sense that a vague attempt is made at 'Definition': '*When someone banks something, they are either depositing it in a financial institution or relying on it for future use*'. Much more convoluted than what COBUILD does: '*If you bank money, you pay it into a bank*'.³

Plus, the first example for this verbal use of *bank* follows the pattern [VERB + in] and the second [VERB + on], so they don't really belong there, while COBUILD itself actually lists [VERB + with] as a verbal sense in itself: '*If you bank with a particular bank, you have an account with that bank*'.⁴



The screenshot shows the ChatGPT-3.5 interface for the lemma 'bank'. At the top left is the OpenAI logo, followed by the word 'bank' and its phonetic transcription '[bæŋk]'. To the right are icons for copy, like, and comment. Below this, the frequency is listed as '*****'. The main content is organized into sections: '1. Noun (Financial Institution)', 'Morphological Forms', 'Sense', 'Usage Pattern', 'Definition', 'Translation Equivalents in Portuguese', 'Example 1 in English', 'Translation 1 in Portuguese', 'Example 2 in English', 'Translation 2 in Portuguese', 'Pronunciation', 'Synonyms', and 'Antonyms'.

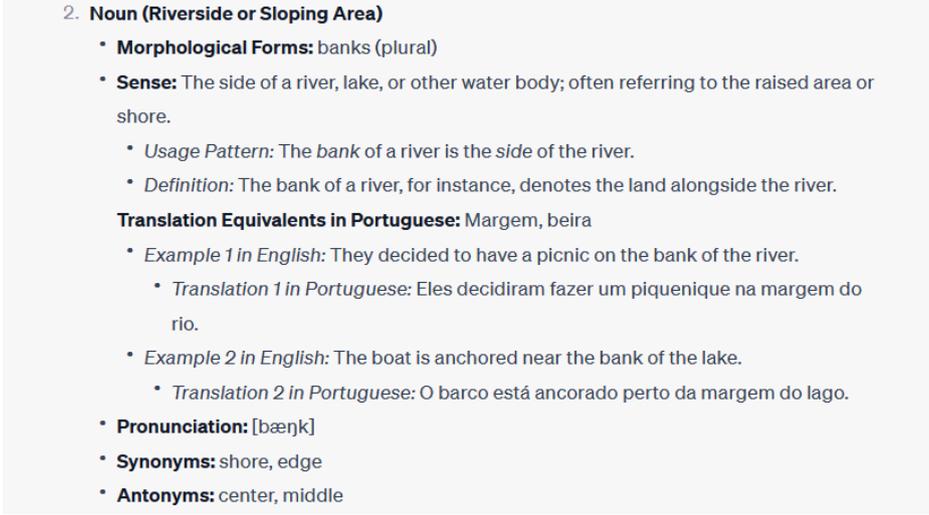
bank [bæŋk]

Frequency: *****

1. **Noun (Financial Institution)**

- **Morphological Forms:** banks (plural)
- **Sense:** A financial institution where individuals and businesses can deposit money, obtain loans, and conduct various financial transactions.
 - *Usage Pattern:* A bank is a place where individuals and businesses can deposit money, obtain loans, and conduct financial transactions.
 - *Definition:* A bank is an establishment where people and companies can place their money, borrow funds, and manage various financial dealings.
- **Translation Equivalents in Portuguese:** Banco
 - *Example 1 in English:* I need to visit the bank to withdraw some cash.
 - *Translation 1 in Portuguese:* Preciso ir ao banco para sacar algum dinheiro.
 - *Example 2 in English:* The bank offers various financial services, including savings accounts and mortgages.
 - *Translation 2 in Portuguese:* O banco oferece diversos serviços financeiros, incluindo contas de poupança e hipotecas.
- **Pronunciation:** [bæŋk]
- **Synonyms:** financial institution, credit union
- **Antonyms:** customer, borrower

Figure 7a: ChatGPT-3.5 output for the English lemma *bank*, # 1, for the prompt seen in Figure 6



The screenshot shows the ChatGPT-3.5 interface for the lemma 'bank'. It continues from the previous section with '2. Noun (Riverside or Sloping Area)'. It follows the same structure as the first section, including 'Morphological Forms', 'Sense', 'Usage Pattern', 'Definition', 'Translation Equivalents in Portuguese', 'Example 1 in English', 'Translation 1 in Portuguese', 'Example 2 in English', 'Translation 2 in Portuguese', 'Pronunciation', 'Synonyms', and 'Antonyms'.

2. **Noun (Riverside or Sloping Area)**

- **Morphological Forms:** banks (plural)
- **Sense:** The side of a river, lake, or other water body; often referring to the raised area or shore.
 - *Usage Pattern:* The bank of a river is the side of the river.
 - *Definition:* The bank of a river, for instance, denotes the land alongside the river.
- **Translation Equivalents in Portuguese:** Margem, beira
 - *Example 1 in English:* They decided to have a picnic on the bank of the river.
 - *Translation 1 in Portuguese:* Eles decidiram fazer um piquenique na margem do rio.
 - *Example 2 in English:* The boat is anchored near the bank of the lake.
 - *Translation 2 in Portuguese:* O barco está ancorado perto da margem do lago.
- **Pronunciation:** [bæŋk]
- **Synonyms:** shore, edge
- **Antonyms:** center, middle

Figure 7b: ChatGPT-3.5 output for the English lemma *bank*, # 2, for the prompt seen in Figure 6

3. Verb

- **Morphological Forms:** banks (3rd person singular present), banked (past tense), banking (present participle), banks (3rd person singular present)
- **Sense:** To deposit or store something in a bank, or to rely on or store something for future use.
 - *Usage Pattern:* Someone banks something in a bank, or they bank on something for a specific purpose.
 - *Definition:* When someone banks something, they are either depositing it in a financial institution or relying on it for future use.
- **Translation Equivalents in Portuguese:** Depositar, guardar, contar com
 - *Example 1 in English:* She banks her savings in a local bank.
 - *Translation 1 in Portuguese:* Ela deposita suas economias em um banco local.
 - *Example 2 in English:* He's banking on winning the competition with his skills.
 - *Translation 2 in Portuguese:* Ele está contando com sua vitória na competição com suas habilidades.
- **Pronunciation:** [bæŋk]
- **Synonyms:** deposit, rely on
- **Antonyms:** withdraw, disbelieve

Figure 7c: ChatGPT-3.5 output for the English lemma *bank*, # 3, for the prompt seen in Figure 6

Furthermore, given that a request was made for an English to Portuguese dictionary, there is also still too much in English. The next prompt became: '*Right, but what happened to the Portuguese? It is supposed to be a bilingual dictionary!*'. After that ChatGPT continued with presenting two versions, asking '*Which response do you prefer?*'. Part of the two versions is shown in Figure 8.

In the version on the left of Figure 8, the 'Definition' is now in Portuguese only, while in the version on the right of Figure 8, the synonyms and antonyms are in Portuguese only. Both additions are fine, but it had been better if both options were merged. Conclusion: As was the case for the monolingual test, it is clear that also for this direction of a bilingual dictionary, ChatGPT-3.5 is not up to the task, as it is unable to compile proper COBUILD-style entries for an English–Portuguese learners' dictionary.

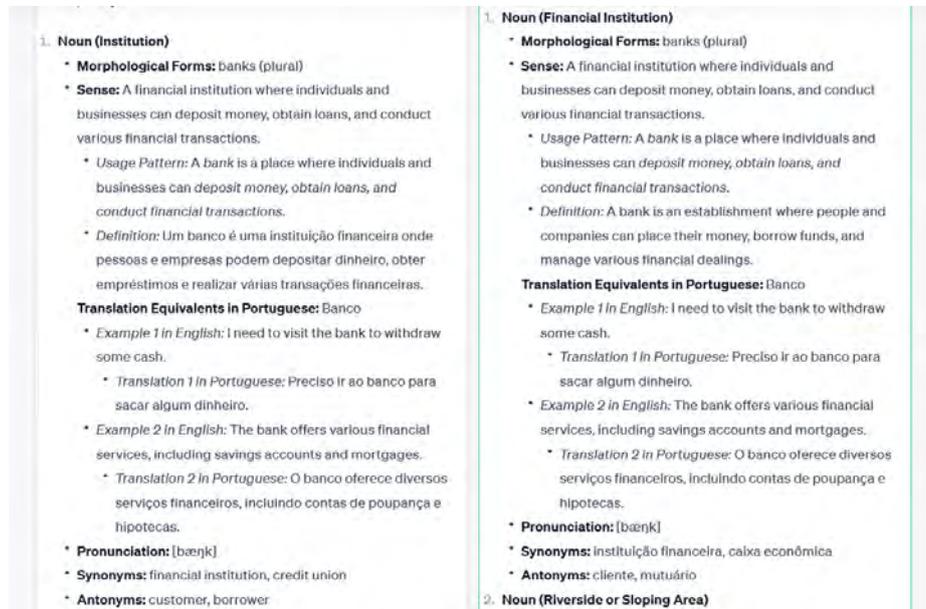


Figure 8: Part of the ChatGPT-3.5 output for the prompt seen in Figure 6, plus a request for more Portuguese, in two versions (left vs. right)

As a final test, the reverse was tried, producing a COBUILD-style Portuguese–English learners' dictionary, with the prompt shown in Figure 9, and the reply as seen in Figures 10a, b and c.



Dê-me um verbete de dicionário bilingue português-ínglês para a palavra "banco". Comece com as informações de frequência na parte superior. Cada sentido deve estar em um bloco numerado. Em seguida, cada bloco começa com a classe gramatical e as formas morfológicas do respectivo sentido. Em seguida, há uma definição do sentido em português e a tradução dessa definição para o inglês. Observe que, para definir, você deve usar o estilo de definição de frase completa, que consiste em duas cláusulas: a primeira cláusula ilustra o padrão de uso, a segunda cláusula é uma paráfrase do significado. Liste também um ou mais equivalentes de tradução em inglês para cada sentido. Adicione exemplos em português com suas traduções em inglês que ilustrem tanto o uso quanto o significado de cada sentido específico em português. Para as frases de exemplo, certifique-se de usar diferentes estruturas de frases, referindo-se a diferentes pessoas; refira-se a situações passadas, presentes e futuras; varie frases de exemplo longas e curtas; e inclua outras elaborações para que você, caro ChatGPT, produza material de dicionário de alta qualidade para mim. Inclua também uma pronúncia do verbete usando o IPA e inclua uma lista de sinônimos e antônimos nos sentidos em que essas listas forem relevantes.

Figure 9: Prompt to obtain a COBUILD-style entry for a bilingual Portuguese–English learners' dictionary, here for the lemma *banco* 'bank'



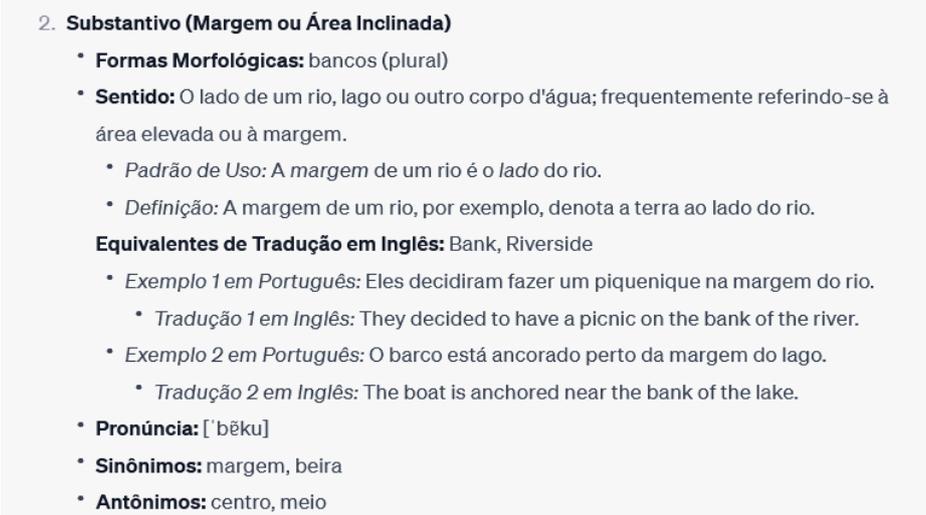
 **banco** [ˈbẽku] 📄 🍌 🗨️

Frequência: *****

1. Substantivo (Instituição Financeira)

- **Formas Morfológicas:** bancos (plural)
- **Sentido:** Uma instituição financeira onde pessoas e empresas podem depositar dinheiro, obter empréstimos e realizar várias transações financeiras.
 - *Padrão de Uso:* Um banco é um local onde pessoas e empresas podem depositar dinheiro, obter empréstimos e realizar transações financeiras.
 - *Definição:* Um banco é um estabelecimento onde pessoas e empresas podem colocar seu dinheiro, obter empréstimos e gerenciar várias transações financeiras.
- **Equivalentes de Tradução em Inglês:** Bank
 - *Exemplo 1 em Português:* Preciso ir ao banco para sacar algum dinheiro.
 - *Tradução 1 em Inglês:* I need to go to the bank to withdraw some money.
 - *Exemplo 2 em Português:* O banco oferece diversos serviços financeiros, incluindo contas de poupança e hipotecas.
 - *Tradução 2 em Inglês:* The bank offers various financial services, including savings accounts and mortgages.
- **Pronúncia:** [ˈbẽku]
- **Sinônimos:** instituição financeira, agência bancária
- **Antônimos:** cliente, mutuário

Figure 10a: ChatGPT-3.5 output for the Portuguese lemma *banco*, # 1, for the prompt seen in Figure 9



2. Substantivo (Margem ou Área Inclinada)

- **Formas Morfológicas:** bancos (plural)
- **Sentido:** O lado de um rio, lago ou outro corpo d'água; frequentemente referindo-se à área elevada ou à margem.
 - *Padrão de Uso:* A margem de um rio é o lado do rio.
 - *Definição:* A margem de um rio, por exemplo, denota a terra ao lado do rio.
- **Equivalentes de Tradução em Inglês:** Bank, Riverside
 - *Exemplo 1 em Português:* Eles decidiram fazer um piquenique na margem do rio.
 - *Tradução 1 em Inglês:* They decided to have a picnic on the bank of the river.
 - *Exemplo 2 em Português:* O barco está ancorado perto da margem do lago.
 - *Tradução 2 em Inglês:* The boat is anchored near the bank of the lake.
- **Pronúncia:** [ˈbẽku]
- **Sinônimos:** margem, beira
- **Antônimos:** centro, meio

Figure 10b: ChatGPT-3.5 output for the Portuguese lemma *banco*, # 2, for the prompt seen in Figure 9

3. Verbo

- **Formas Morfológicas:** banco (3ª pessoa do singular do presente), bancou (pretérito), bancando (particípio presente), banco (3ª pessoa do singular do presente)
- **Sentido:** Depositar ou armazenar algo em um banco, ou confiar ou armazenar algo para uso futuro.
 - *Padrão de Uso:* Alguém *bancou* algo em um *banco*, ou eles *bancam* algo para um propósito específico.
 - *Definição:* Quando alguém deposita algo em um banco, está armazenando em uma instituição financeira ou confiando em algo para uso futuro.
- **Equivalentes de Tradução em Inglês:** Deposit, Rely on
 - *Exemplo 1 em Português:* Ela deposita suas economias em um banco local.
 - *Tradução 1 em Inglês:* She banks her savings in a local bank.
 - *Exemplo 2 em Português:* Ele está contando com sua vitória na competição com suas habilidades.
 - *Tradução 2 em Inglês:* He's banking on winning the competition with his skills.
- **Pronúncia:** [ˈbẽku]
- **Sinônimos:** depositar, confiar
- **Antônimos:** sacar, desacreditar

Figure 10c: ChatGPT-3.5 output for the Portuguese lemma *banco*, # 3, for the prompt seen in Figure 9

When the output seen in Figures 10a, b and c was presented to two professors in Brazil, one a lexicographer for Portuguese (Regiani Aparecida Santos Zacarias), the other a linguist for Portuguese (who prefers to remain anonymous), they thought it all looked rather good at first. When they were asked (too) many questions, their opinion changed. While the first sense (Figure 10a), for the nominal use as a financial institution, is acceptable, the second sense (Figure 10b), for the nominal use as a sloping area, and third sense (Figure 10c), as a verb, are riddled with errors. The second block deals with the meaning 'bank of a river'. Even though this is supposed to be a sense for the lemma *banco*, the word used under 'Sense', 'Usage Pattern' and 'Definition' as well as in the two examples is a different word entirely: *margem*. At the same time, under the synonym list for *banco* one finds *margem* 'margin; riverbank; shore' and *beira* 'edge; riverbank; precipice verge'. Ironically, and as can already be deduced, only the synonyms presented refer to the bank of a river, using *banco* for that is simply not possible.

The third sense of '*banco*' supposedly presents the verbal use of this lemma. However, the sense 'to bank money' simply does not exist with the lemma *banco* in Portuguese. Now, ChatGPT not only 'created' this sense, it went as far as

inventing conjugations for *banco* to then use these conjugations as well (**bancou*, **bancam*), then continues with a convincing COBUILD-style full-sentence definition but forgets to use the lemma and instead uses the (correct) verbs *depositar* 'deposit' and *confiar* 'rely on', and finally does the same when providing the first example. Likewise for the phrasal verb 'count on', so [VERB + on], which also does not use *banco* in Portuguese, but rather *contar com* as seen in the second example of Figure 10c. When ChatGPT offers its list of synonyms, it merely repeats what it had used all along in the microstructure already. Note that there is a verb *bancar* 'fund, finance' in Portuguese, for which the morphological forms are *banca*, *bancou*, *bancando*, *banca*.⁵ Conclusion: As was the case for the other dictionary direction, ChatGPT-3.5 is once more not up to the task, as it is fully unable to compile proper COBUILD-style entries for a Portuguese–English learners' dictionary.

So, what is really going on, and what is ChatGPT-3.5 doing for bilingual lexicography? Well, the assumption is that one is actually looking at English lexical analysis through the eyes of a non-English language, here Portuguese. Reformulated, the hypothesis is that everything one sees there was created internally based on English (given the strong bias in the training data) and was merely mapped onto Portuguese since the prompts asked to go bilingual. If a further analogy is allowed: Given the overwhelming amount of English in an LLM, a model like ChatGPT-3.5 is simply incapable to 'think' in any other language, let alone analyse it linguistically in any proper way.

2.3 Low-visible Bantu

Back in 2023–2024, with the certainty that an LLM like ChatGPT is 'fantastic' for English, and the knowledge that it is basically 'fake' for bilingual lexicography, does that mean that it simply 'fails' for exotic languages? During a talk at the University of Missouri (De Schryver 2024b) various attempts were made at having ChatGPT compile proper dictionary articles for a variety of Bantu languages, amongst others Swahili (100 million speakers; spoken in more than 14 countries), Xhosa (19 million speakers; spoken in 2 countries) and Lusoga (4.5 million speakers; spoken in part of one country).

2.3.1 Region-visible Swahili

First, and again using ChatGPT-3.5, the chatbot was questioned about some of the basics of Swahili, to see whether it gets at least the essentials right for the language. The opening prompt and answer are displayed in Figure 11.

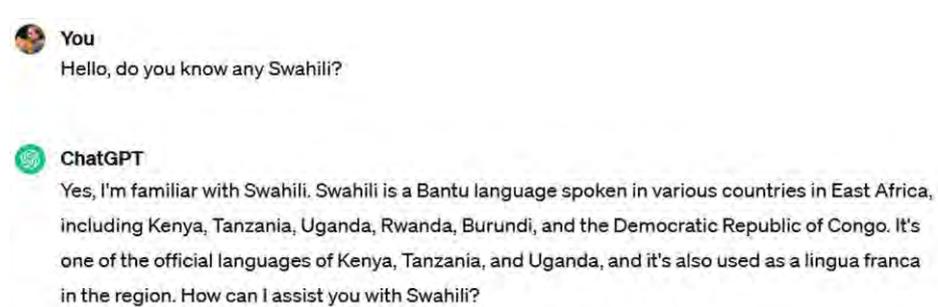


Figure 11: ChatGPT-3.5 opening prompt and answer on the Swahili language

While the answer is only approximately correct (apart from the army, no one uses Swahili in Uganda), the next question and answer immediately delve into lexicography for Swahili, as seen in Figure 12.

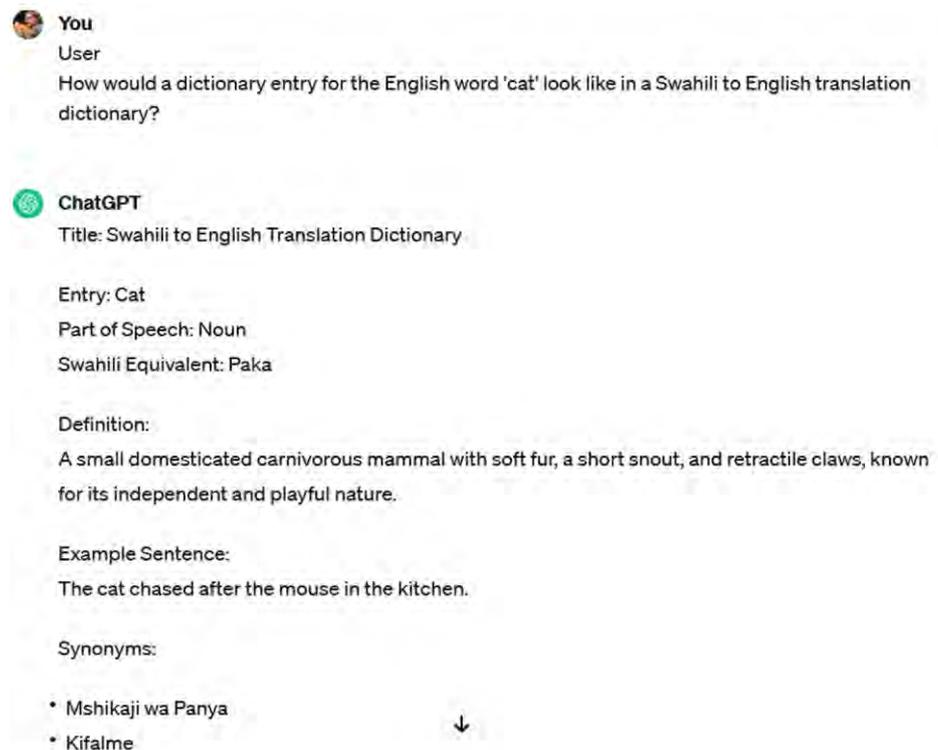


Figure 12: ChatGPT-3.5 prompt and output for an entry in a Swahili–English translation dictionary, here for *cat*

As may be seen from Figure 12, right from the moment one tries out lexicography in Swahili, the chatbot already gets the dictionary direction wrong. The

question was to go from Swahili to English, but the reply came in the reverse dictionary direction, with, strangely, only synonyms (of the English word *cat*) in Swahili. Now, in passing, it would be welcome if a chatbot could start taking over from an online Swahili dictionary, such as the unidirectional Swahili to English dictionary at africanlanguages.com⁶, having been ranked by Google⁷ as the top Swahili dictionary for two decades now, even though the year before its launch, there were already 20 online dictionaries for Swahili (De Schryver 2003: 12). See Addendum B.

In any case, the direction error was pointed out, and ChatGPT then offered the revised material displayed in Figure 13.

As may be seen from Figure 13, ChatGPT basically swapped the lemma and its translation equivalent around, and (correctly) translated the English to Swahili for the definition and example sentence. All of this, together with the POS, is correct, while the first synonym is original: *mshikaji wa panya* 'mouse catcher', while the second is not clear: *kifalme* 'princess'.

Given that this reply (mostly) made sense, the next prompt asked to see more *uses* and thus more *meanings*, see Figure 14. For more on the core concept of 'mapping *meaning* onto *use*' see Hanks (2002).

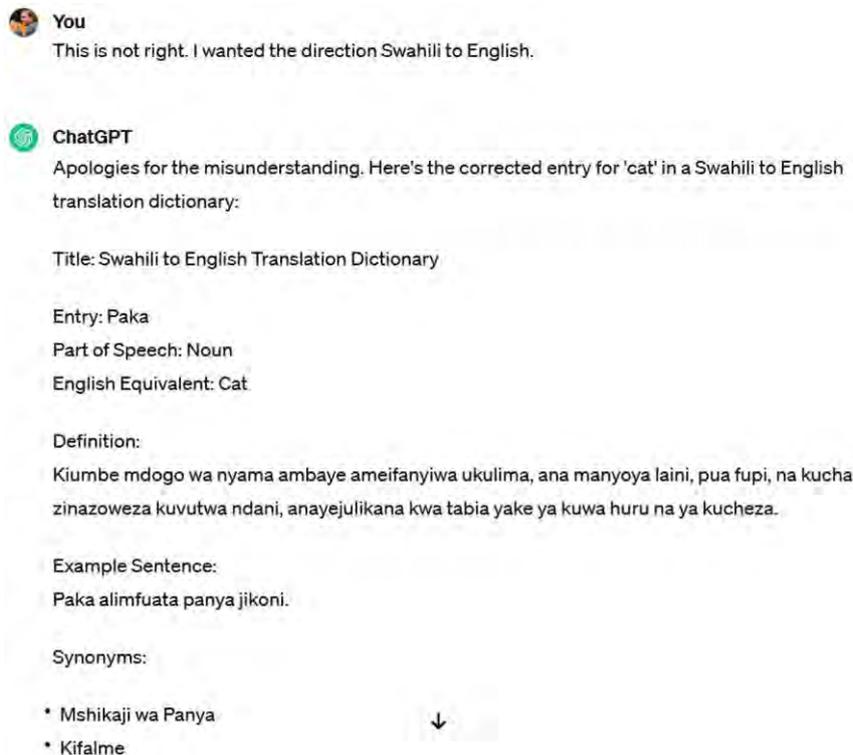


Figure 13: Additional ChatGPT-3.5 prompt and output for an entry in a Swahili–English translation dictionary, here for *paka* 'cat'



You

OK, makes sense. Does it have other uses and thus also other meanings?



ChatGPT

Yes, the word "paka" in Swahili primarily refers to a domestic cat, but it can also be used in other contexts or idiomatic expressions, giving rise to additional meanings or connotations. Here are a few examples:

- 1. Verb Conjugation:** In Swahili, "paka" can also be used as a verb, meaning "to carry" or "to wear" when describing actions related to carrying something or wearing clothes. For example:
 - Nina paka begi langu. (I am carrying my bag.)
 - Anapaka kofia nyekundu. (He/she is wearing a red hat.)
- 2. Idiomatic Expressions:** The word "paka" might be used in idiomatic expressions with meanings that extend beyond the literal translation. For instance:
 - Kuwa na paka mbili ni shida. (Literal translation: Having two cats is a problem.)
 - This expression might mean that having too many of something can lead to complications or difficulties.
- 3. Metaphorical Use:** Sometimes "paka" can be used metaphorically to describe someone's behavior or characteristics, often implying slyness or cunning similar to the characteristics attributed to cats in various cultures.

Overall, while "paka" primarily means "cat" in Swahili, its usage can extend to various contexts and idiomatic expressions, allowing for different interpretations and meanings.

Figure 14: Second additional ChatGPT-3.5 prompt and output for an entry in a Swahili–English translation dictionary, here for *paka* 'cat'

As may be seen from Figure 14, this reply went horribly wrong. Instead of seeing an entry with multiple (numbered) senses, all the earlier information has been dispensed with. As for the three new numbered paragraphs, the first makes no sense at all, as it mixes up the noun *paka* 'cat' with the verb *-paka* 'smear, spread, apply' — which is anything but another sense of the noun *cat*; the second paragraph is an idiomatic expression with the noun *cat* — which is good; while the third paragraph talks about metaphorical uses in English, but there is not a single usage example in either language — so not useful. The wrap-up at the end is correct, but does not add anything new.

Courtesy obliges, one needs to end with the verb of communication *-pendekeza* 'recommend'. Figure 15 shows the usual prompt (which was also used for Portuguese in Figures 4 and 9), but now translated into Swahili, to obtain a COBUILD-style entry for a monolingual Swahili learners' dictionary. For the translation, here and further on, see Addendum C.

 **You**
Nipe ingizo la kamusi ya Kiswahili-Kiingereza kwa lugha mbili kwa neno '-pendekeza'. Anza na maelezo ya marudio hapo juu. Kila maana inapaswa kuwa katika block iliyohesabiwa. Kisha kila kizuizi huanza na sehemu ya hotuba na maumbo ya kimofolojia ya maana husika. Hii inafuatwa na ufafanuzi wa maana katika Kiswahili, na tafsiri ya ufafanuzi huo kwa Kiingereza. Kumbuka kwamba ili kufafanua, inabidi utumie mtindo wa kufafanua sentensi kamili, ambao una vishazi viwili: kifungu cha kwanza kinaonyesha muundo wa matumizi, kifungu cha pili ni kifungu cha maana. Pia orodhesha tafsiri moja au zaidi zinazolingana katika Kiingereza katika kila maana. Ongeza mifano ya Kiswahili pamoja na tafsiri zao za Kiingereza zinazoonyesha matumizi na maana ya kila maana ya Kiswahili. Kwa sentensi za mfano, hakikisha unatumia miundo tofauti ya sentensi, ukirejelea watu tofauti; rejea hali zilizopita, za sasa, na zijazo; tofautisha sentensi ndefu na fupi za mifano; na ujumuishe maelezo mengine ili wewe, mpendwa ChatGPT, unitayarishie nyenzo za ubora wa juu za kamusi. Pia jumuishe matamshi ya ingizo kwa kutumia IPA, na ujumuishe orodha ya visawe na vinyume kwenye hisi ambapo orodha kama hizo zinafaa.

Figure 15: ChatGPT-3.5 prompt to obtain a COBUILD-style entry for a monolingual Swahili learners' dictionary, here for the verb *-pendekeza* 'recommend'

The reply by ChatGPT looked decidedly weird, as may be seen from Figure 16 and the explanation that follows.

 **ChatGPT**
-pendekeza

Marudio: -pendekeza (Kitenzi)

1. Kitenzi - Tendeka

Maana: Kutoa maoni au mapendekezo ya kufanya kitu fulani kwa sababu unaamini kuwa ni bora au inafaa kufanywa.

Tafsiri: recommend

Visawe: kupendekeza, kushauri

Vinyume: kukataa, kupinga

Mifano:

- *Daktari alimpendekeza mgonjwa apumzike zaidi.*
(The doctor recommended that the patient rest more.)
- *Mwalimu alitupendekeza tufanye utafiti wa ziada kwa ajili ya mtihani ujao.*
(The teacher recommended that we do extra research for the upcoming exam.)

Matamshi: /pende' keza/

Visawe na Vinyume:

- Visawe: kupendekeza, kushauri
- Vinyume: kukataa, kupinga



Figure 16: ChatGPT-3.5 output for the Swahili lemma *-pendekeza* 'recommend', for the prompt seen in Figure 15

From Figure 16, it is quickly clear that what ChatGPT-3.5 produces is of little value: The lemma is repeated twice, the part of speech (*kitenzi* 'verb') is repeated twice, the synonym and antonym slots are repeated twice, and the first so-called synonym simply takes the lemma and puts it into class 15, which is nothing special, as all verbs in any Bantu language belong to class 15, and thus start with *ku-*. The definition does not follow the COBUILD-style, and even so, turns the verb into the noun *mapendekezo* 'recommendation' to explain the verb *-pendekeza* 'recommend', which is bad lexicography. And in terms of layout, there is also no reason for the number 1, at the first and only *kitenzi* 'verb' block. Conclusion: ChatGPT-3.5 can simply not handle Swahili lexicography.

2.3.2 Nation-visible Xhosa

Given that Xhosa is a much smaller Bantu language compared to Swahili, basically only spoken in South Africa and recognized as a minority language in Zimbabwe, the approach this time was to try to make it super-simple, by merely asking for a small number of random short dictionary entries for Xhosa, as seen in Figure 17.



8. "Hamba kahle" - Goodbye
9. "Indlu" - House
10. "Inkosana" - Child
11. "Ibhokisi" - Box
12. "Iziko" - Museum
13. "Indawo" - Place
14. "Umntu" - Person
15. "Umntwana" - Child
16. "Umlilo" - Fire
17. "Inja" - Dog
18. "Inyoni" - Bird
19. "Igqirha" - Doctor
20. "Isango" - Door



Figure 17: ChatGPT-3.5 prompt and reply to obtain a small number of random short dictionary entries for Xhosa

Figure 17 hardly displays a list of dictionary entries, only just a random vocabulary list. Unfortunately, it is abundantly clear that ChatGPT cannot even pick 20 frequent words from Xhosa, as item number 18, *inyoni* 'bird', is from the neighbouring and bigger language Zulu (27 million speakers; spoken in at least 6 countries), see Figure 18.

inyoni ** *noun* 9/10 (*pl. izinyoni*) ■ bird
♦ Silalel' izinyoni zitshiloza. *We listen to the birds chirping.*
▶ **kwelenyoni** ■ topmost part of a tree
♦ Wathi lapho esephezulu kwelenyoni, wase elula isandla ukuze abambe igatsha elithele kakhulu izithelo. *When he was at the top of the tree, he stretched out his hand to grab a branch that was full of fruit.*

Figure 18: Dictionary entry for *inyoni* 'bird' in the *Oxford Zulu School Dictionary* (De Schryver 2010: 96)

Rather, the word for bird in Xhosa is *intaka*, see Figure 19.

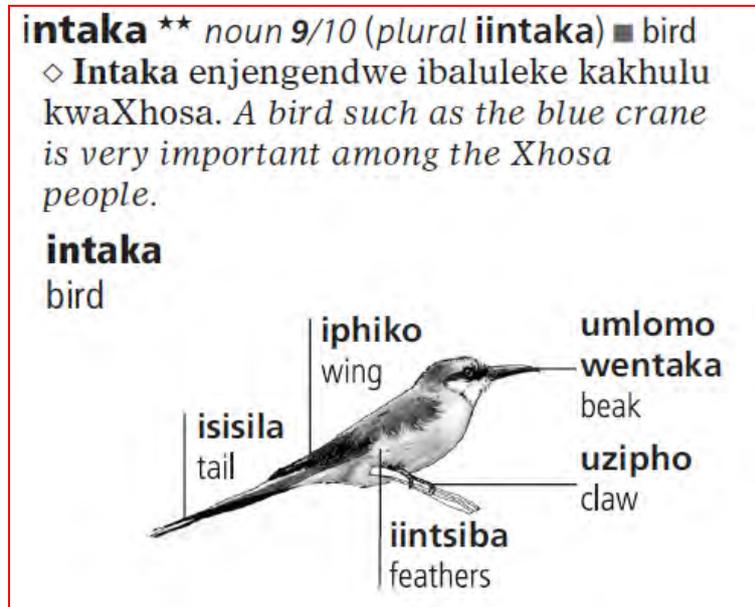


Figure 19: Dictionary entry for *intaka* 'bird' in the *Oxford Xhosa School Dictionary* (De Schryver and Reynolds 2014: 89)

Even several translations are debatable: item 10, *inkosana*, is translated as 'child', but see Figure 20, where it is prince.

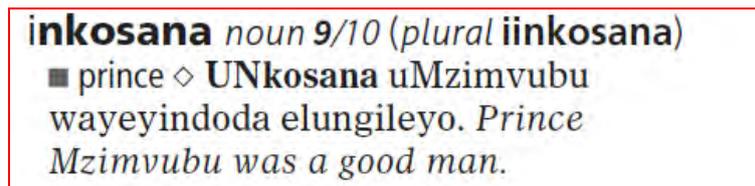


Figure 20: Dictionary entry for *inkosana* 'prince' in the *Oxford Xhosa School Dictionary* (De Schryver and Reynolds 2014: 87)

Item 12, *iziko*, is translated as 'museum', but see Figure 21, where it is '1. fireplace; 2. centre; institution'.

iziko ** *noun* 5/6 (plural **amaziko**)
1 ■ fireplace ◇ Lumka, **iziko** lisheshushu. *Be careful, the fireplace is still hot.* **2** ■ centre; institution ◇ **IZiko** leeLwimi likhuthaza zonke iilwimi. *The Centre for Languages encourages the use of all languages.*

Figure 21: Dictionary entry for *iziko* '1. fireplace; 2. centre; institution' in the *Oxford Xhosa School Dictionary* (De Schryver and Reynolds 2014: 119)

While the translation for museum is actually a loanword, namely *imyuziyam*, see Figure 22.

museum * (say myoo-zee-uhm) *noun* (plural **museums**) ■ imyuziyam ◇ In this **museum** you can see many interesting things from Xhosa history. *Kule myuziyam uya kubona izinto ezininzi ezinika umdla ezivela kwimbali yamaXhosa.*

Figure 22: Dictionary entry for *museum* in the *Oxford Xhosa School Dictionary* (De Schryver and Reynolds 2014: 418)

Therefore, and as an overall conclusion so far, with ChatGPT-3.5 English not only does much better than other big languages (like Portuguese), but small languages (like Zulu) take over even smaller ones (like Xhosa).

2.3.3 Local-visible Lusoga

The next level, and still within Bantu, concerns what happens to even smaller languages of the primarily oral type, i.e. Bantu languages for which there is hardly any written documentation, no matter the actual number of sometimes impressively high number of speakers. Lusoga is such a case, spoken by up to 4.5 million speakers in the southeast of Uganda. The prompt was kept even more basic than any of the previous ones, just asking for the top frequent words in the language, as displayed in Figure 23.

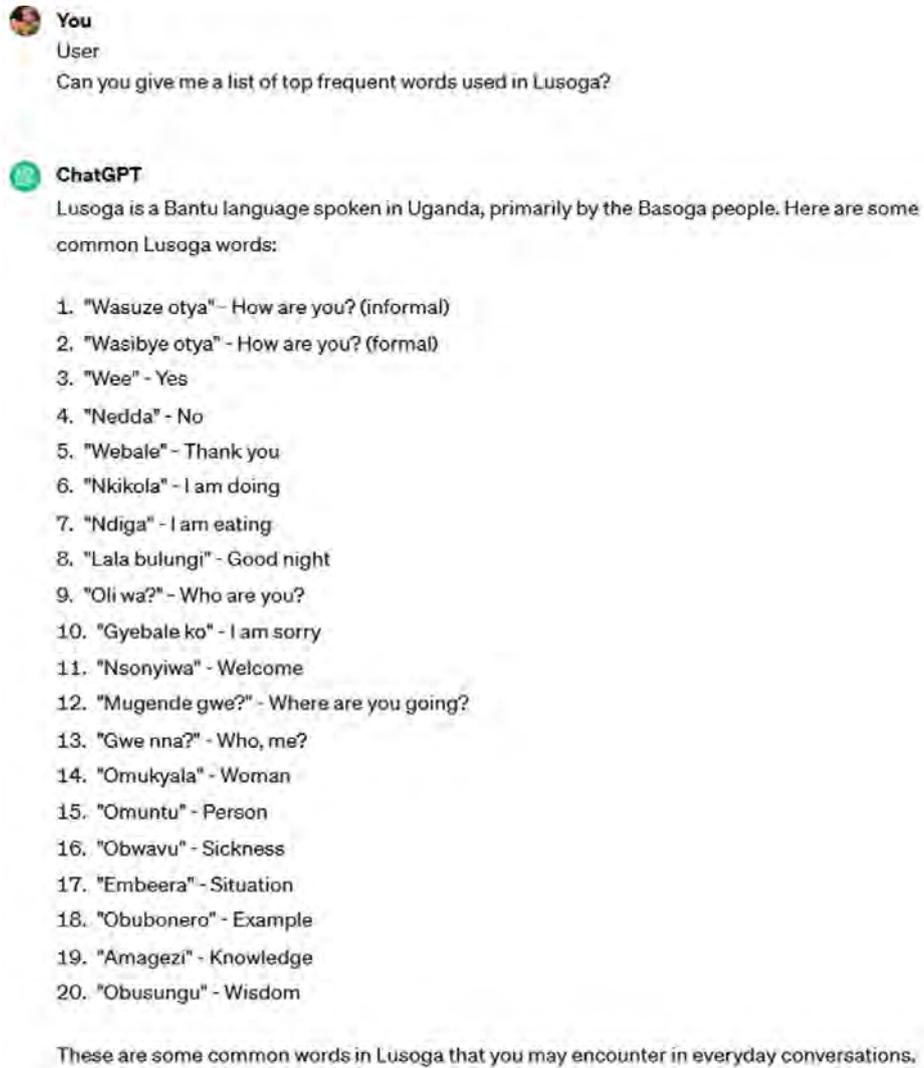


Figure 23: ChatGPT-3.5 prompt and reply to obtain a list of the frequent words in Lusoga

Firstly, this list has moved away from a list of words, to mostly a list of phrases. Secondly, an analysis of the ChatGPT-3.5 output by Minah Nabirye, a native speaker of Lusoga, reveals that only 7 out of the 20 are correct, with two of those correct ones actually only in the neighbouring and much bigger language Luganda (20 million speakers; spoken in Uganda immediately to the west of Lusoga, as well as in Rwanda). See Table 1 for the details.

Table 1: Analysis of the ChatGPT-3.5 output of so-called 'frequent words in Lusoga'

ChatGPT-3.5	ChatGPT-3.5 translation	Language	Status	Lug(anda)	Lus(oga)	English
1 <i>Wasuze otya</i>	How are you? (informal)	Lug/Lus	Wrong translation	-	-	Good morning
2 <i>Wasibye otya</i>	How are you? (formal)	Lug/Lus	Wrong form	<i>Osiibye otya?</i>	<i>Osiibye otya?</i>	-
3 <i>Wee</i>	Yes	??	Unknown form	-	-	-
4 <i>Nedda</i>	No	Lug	Correct	-	-	-
5 <i>Webale</i>	Thank you	Lug/Lus	Wrong form	<i>Weebale</i>	<i>Weebale</i>	-
6 <i>Nkikola</i>	I am doing	Lug	Correct	-	-	-
7 <i>Ndiga</i>	I am eating	Lug	Wrong translation	-	-	Sheep
8 <i>Lala bulungi</i>	Good night	Lug/Lus	Wrong form	<i>Sula bulungi</i>	<i>Sula bulungi</i>	-
9 <i>Oli wa?</i>	Who are you?	Lug/Lus	Wrong translation	-	-	Where are you?
10 <i>Gyebale ko</i>	I am sorry	Lug	Wrong translation	<i>Gyebaleko</i>	<i>Gyebaleku</i>	Thank you for your work
11 <i>Nsonyiwa</i>	Welcome	Lug	Wrong translation	-	<i>Nsonhiwa</i>	Forgive me
12 <i>Mugende gwe?</i>	Where are you going?	??	Unknown form	<i>Mugende (*gwe)</i>	<i>Mugye (*iwe)</i>	You go(pl)(gwe/iwe = you)
13 <i>Gwe nna?</i>	Who, me?	??	Unknown form	-	-	-
14 <i>Omukyala</i>	Woman	Lug/Lus	Correct	-	-	-
15 <i>Omuntu</i>	Person	Lug/Lus	Correct	-	-	-
16 <i>Obwavu</i>	Sickness	Lug/Lus	Wrong translation	-	-	Poverty
17 <i>Embeera</i>	Situation	Lug/Lus	Correct	-	-	-
18 <i>Obubonero</i>	Example	Lug/Lus	Correct	-	-	-
19 <i>Amagezi</i>	Knowledge	Lug/Lus	Correct	-	-	-
20 <i>Obusungu</i>	Wisdom	Lug/Lus	Wrong translation	-	-	Anger

That means that all in all, there are merely 5 correct Lusoga forms out of 20, or thus one quarter (25%). Conclusion: If three quarters (75%) of the returned material is erroneous on the most basic of the basic lexicographic tasks, there is simply no point in trying anything else with ChatGPT-3.5 for Lusoga, even though it is spoken by a respective number of as many as 4.5 million speakers.

3. What to do next?

The main take-away from this article is that the lexicographic prompt introduced by De Schryver (2023b), being a single prompt with which to compile an entire English-style COBUILD learners' dictionary, may indeed work well for English, but is not good enough for other languages. At least, not using ChatGPT-3.5. The field of GenAI chatbots, however, has moved on since then, so a serious attempt must first be made to see how later out-of-the-box models like GPT-4o or Gemini-2.5 perform the job on other languages than English. Second, an attempt should be made to move from an out-of-the-box model, to one where one starts customising the model. These two aspects are precisely the topics of a follow-up article (see De Schryver 2025).

PS.

In this day and age, and as the secretary of the exam committee for African Studies at UGent where the author of this publication is tasked with checking

all BAs, MAs and PhDs on their correct use of GenAI, the reader needs to be informed that GenAI was *not* used for any of the writings in this article, except for some of the sub-headings. Addendum D, available online (click [here](#)), displays the full chat between the author and OpenAI's GPT-4o in this respect. In itself, it provides an interesting perspective on using GenAI as a writing assistant. Research on using AI as a writing assistant in our field started with the work of Henrik Køhler Simonsen, first published as Leroyer and Simonsen (2019), and from then on, year after year, as Simonsen (2020, 2021, 2022, 2023). Subsequent to the release of ChatGPT, other colleagues followed (Huete-García and Tarp 2024, Tarp and Nomdedeu-Rull 2024).

Reading through the entire chat of Addendum D, available online (click [here](#)), reveals how good GPT-4o has become as a lexicographer; it is like consulting a wide variety of six different dictionaries, going in all sort of directions: (1) magnitudes, scales, sizes, (2) demographic speaker populations, (3) socio-technological reach, visibility, language power, (4) digital inclusion, corpus resources, language equity, (5) localness without grandeur, geographic nuance, judgmental rings, and (6) progression from hyper to local, capturing visibility and rank, balancing clarity, tone and structure. Simply stunning, and it is unlikely that any human lexicographer can do this in the fractions of seconds that GPT-4o spends on each of those levels. For the interested reader, even some screen recordings exist, from which the varied speed of the GenAI replies can be monitored. That said, one of the adjudicators observed all of the following, with which I fully agree:

That's not a fair point. It is, in my opinion overly simplistic and reductive to use generation speed as the sole criterion. What about training time or about the time it takes to critically check the responses? If one argues at this level, one could also point to the questionable use of resources in training and running current LLMs: not only does this speed come at the price of enormous resource consumption (electricity and water), but the language training material was also most likely acquired in violation of copyright laws. When praising the speed of LLM's responses, I believe it is important not to remain silent about these issues. (Anonymous adjudicator for this article, to whom sincere thanks are due.)

Endnotes

1. See <https://www.collinsdictionary.com/dictionary/english/strike>, sense 3.
2. For more on the FSD, see Hanks (1987), Barnbrook (2002), Lew (2023), and Rundell (2024).
3. See <https://www.collinsdictionary.com/dictionary/english/bank>, finance and storage, sense 3.
4. See <https://www.collinsdictionary.com/dictionary/english/bank>, finance and storage, sense 4.
5. See <https://www.definify.com/word/bancar>.
6. See <https://africanlanguages.com/swahili/>.
7. See <https://www.google.com/>.

References

- Barnbrook, G.** 2002. *Defining Language: A Local Grammar of Definition Sentences* (Studies in Corpus Linguistics 11). Amsterdam: John Benjamins.
- Chomsky, N.** 1957. *Syntactic Structures* (Janua Linguarum 4). The Hague: Mouton.
- De Schryver, G.-M.** 2003. Online Dictionaries on the Internet: An Overview for the African Languages. *Lexikos* 13: 1-20.
- De Schryver, G.-M.** 2010. *Oxford Bilingual School Dictionary: Zulu and English / Isichazamazwi Sesikole Esinezilimi Ezimbili: IsiZulu NesiNgesi, Esishicilelwe abakwa-Oxford*. Cape Town: Oxford University Press Southern Africa.
- De Schryver, G.-M.** 2023a. *Contemporary Lexicography: A Case Study in AI [Keynote Lecture]*. Paper presented at the Inaugural Conference of the Association of Lexicography for the Americas: South, Central, Caribbean, and Mexico (Americalex-S), São Paulo, Brazil, 20–25 October 2023. <https://youtu.be/watch?v=yAxzTx7A2LU>
- De Schryver, G.-M.** 2023b. Generative AI and Lexicography: The Current state of the Art Using ChatGPT. *International Journal of Lexicography* 36(4): 355-387 + ten pages of supplementary material online.
- De Schryver, G.-M.** 2024a. *Customising LLMs for Lexicography [Keynote Lecture]*. Paper presented at the 8th International Workshop on Cognitive Aspects of the Lexicon (CogALex-VIII), co-located with the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Turin, Italy, 20 May 2024.
- De Schryver, G.-M.** 2024b. *Generative Artificial Intelligence & Three Lexicographic Fs: FANTASTIC for Monolingual English Dictionaries, FAKE for Translation Dictionaries, FAIL for Exotic Dictionaries [Invited Talk]*. Paper presented at the Department of English, University of Missouri, Columbia, MO, USA, 5 April 2024. https://youtu.be/zUI0GrPrH_0
- De Schryver, G.-M.** 2024c. The Road towards Fine-tuned LLMs for Lexicography. Krek, S. (Ed.). 2024. *Book of Abstracts of the Workshop 'Large Language Models and Lexicography' @ EURALEX 2024, 8 October 2024, Cavtat, Croatia: 6-11*. Ljubljana: ELEXIS Association.
- De Schryver, G.-M.** 2025. Customised GPTs for Lexicography. *Lexikos* 35(2): 397-422.
- De Schryver, G.-M.** forthcoming-a. 100 Scholarly Echoes of Generative AI in Lexicography.
- De Schryver, G.-M.** forthcoming-b. Citing the Future: The Legacy of Patrick Hanks.
- De Schryver, G.-M. and M. Reynolds.** 2014. *Oxford Bilingual School Dictionary: IsiXhosa and English / Oxford isiXhosa-isiNgesi English-isiXhosa Isichazi-magama Sesikolo*. Cape Town: Oxford University Press Southern Africa.
- Fuertes-Olivera, P.A.** 2024. Making Lexicography Sustainable: Using ChatGPT and Reusing Data for Lexicographic Purposes. *Lexikos* 34(1): 123-140.
- Hanks, P.** 1987. Definitions and Explanations. Sinclair, J.M. (Ed.). 1987. *Looking Up: An Account of the COBUILD Project in Lexical Computing and the Development of the Collins COBUILD English Language Dictionary: 116-136*. London: Collins ELT.
- Hanks, P.** 2000. Do Word Meanings Exist? *Computers and the Humanities* 34(1–2): 205-215.
- Hanks, P.** 2002. Mapping Meaning onto Use. Corréard, M.-H. (Ed.). 2002. *Lexicography and Natural Language Processing. A Festschrift in Honour of B.T.S. Atkins: 156-198*. s.l.: Euralex.
- Hanks, P.** 2013. *Lexical Analysis: Norms and Exploitations*. Cambridge, MA: The MIT Press.

- Huete-García, Á. and S. Tarp.** 2024. Training an AI-based Writing Assistant for Spanish Learners: The Usefulness of Chatbots and the Indispensability of Human-assisted Intelligence. *Lexikos* 34: 21-40.
- Jakubíček, M. and M. Rundell.** 2023. The End of Lexicography? Can ChatGPT Outperform Current Tools for Post-editing Lexicography? Medved', M., M. Měchura, C. Tiberius, I. Kosem, J. Kallas, M. Jakubíček and S. Krek (Eds.). 2023. *Proceedings of the eLex 2023 Conference: Electronic Lexicography in the 21st Century*: 518-533. Brno: Lexical Computing.
- Leroyer, P. and H.K. Simonsen.** 2019. Google Translate: trussel eller redning for oversættelsesordbøger? [Google Translate: Threat or Salvation for Translation Dictionaries?]. *LexicoNordica* 26: 95-115.
- Lew, R.** 2023. ChatGPT as a COBUILD Lexicographer. *Humanities and Social Sciences Communications* 10(704): 10 pp.
- Lew, R.** 2024. Dictionaries and Lexicography in the AI Era. *Humanities and Social Sciences Communications* 11(426): 8 pp.
- Lew, R., B. Ptasznik and S. Wolfer.** 2024. The Effectiveness of ChatGPT as a Lexical Tool for English, Compared with a Bilingual Dictionary and a Monolingual Learner's Dictionary. *Humanities and Social Sciences Communications* 11(1324): 10 pp.
- McKean, E. and W. Fitzgerald.** 2023. The ROI of AI in Lexicography [Keynote Lecture]. 2023. *Proceedings of the 16th International Conference of the Asian Association for Lexicography: "Lexicography, Artificial Intelligence, and Dictionary Users"*: 18-27. Seoul: Yonsei University.
- Nichols, W.** 2023. Invisible Lexicographers, AI, and the Future of the Dictionary [Keynote Lecture]. Medved', M., M. Měchura, C. Tiberius, I. Kosem, J. Kallas, M. Jakubíček and S. Krek (Eds.). 2023. *Book of Abstracts of the eLex 2023 Conference: Electronic Lexicography in the 21st Century*: 3-3. Brno: Lexical Computing.
- Pavydytė, K.** 2024. *ChatGPT generuojami iliustraciniai laukai: turinio ir kokybės analizė bei palyginimas su BLKŽ [Illustrative Sentences Generated by ChatGPT: The Analysis of Its Content, Quality, and Comparison with the Dictionary of the Standard Lithuanian Language]*. MA Dissertation. Vilnius: Vilnius University.
- Ptasznik, B., S. Wolfer and R. Lew.** 2024. A Learners' Dictionary Versus ChatGPT in Receptive and Productive Lexical Tasks. *International Journal of Lexicography* 37(3): 322-336.
- Rees, G.P. and R. Lew.** 2024. The Effectiveness of OpenAI GPT-generated Definitions Versus Definitions from an English Learners' Dictionary in a Lexically orientated Reading Task. *International Journal of Lexicography* 37(1): 50-74.
- Rundell, M.** 2024. Obituary: Patrick Wyndham Hanks. *International Journal of Lexicography* 37(2): 135-142.
- Simonsen, H.K.** 2020. Augmented Writing: Nye muligheder og nye teorier [Augmented Writing: New Possibilities and New Theories]. Sandström, C., U.-M. Forsberg, C. af Hällström-Reijonen, M. Lehtonen and K. Ruppel (Eds.). 2020. *Nordiska studier i lexikografi 15, Rapport från 15 konferensen om lexikografi i Norden, Helsingfors, 4-7 juni 2019 [Nordic Studies in Lexicography 15, Report from the 15th Nordic Lexicography Conference, Helsinki, 4-7 June 2019]*: 307-315. Helsinki: Nordic Association for Lexicography.
- Simonsen, H.K.** 2021. AI Writers in Language Learning. Chang, M., N.-S. Chen, D.G. Sampson and A. Tlili (Eds.). 2021. *Proceedings of the 21st International Conference on Advanced Learning Technologies (ICALT 2021), Online, 12-15 July 2021*: 238-240. Los Alamitos, CA: IEEE.

- Simonsen, H.K.** 2022. AI Text Generators and Text Producers. Chang, M., N.-S. Chen, M. Dascalu, D.G. Sampson, A. Tlili and S. Trausan-Matu (Eds.). 2022. *Proceedings of the 22nd International Conference on Advanced Learning Technologies (ICALT 2022), Bucharest, Romania, 1–4 July 2022*: 218-220. Los Alamitos, CA: IEEE.
- Simonsen, H.K.** 2023. Bilingvale ordbøger vs. AI-skriveassistenter? [Bilingual Dictionaries vs. AI Writing Assistants?]. *LexicoNordica* 30: 117-136.
- Tarp, S. and A. Nomdedeu-Rull.** 2024. Who Has the Last Word? Lessons from Using ChatGPT to Develop an AI-based Spanish Writing Assistant. *Círculo de Lingüística Aplicada a la Comunicación [Circle of Linguistics Applied to Communication]* 97: 309-321.
- Zhong, A.** 2025. Prompts for Language Learners: A Practical Guide to Using DeepSeek as a Dictionary. *Lexikos* 35(1): 274-285.

Addendum A: Using ChatGPT-3.5 — For Portuguese Lexicography [translations into English]

- Translation of Figure 1:

Please give me a COBUILD-style monolingual dictionary entry for 'strike'. Start with the frequency information at the top. Each sense should be in a numbered block. Each block then starts with the part of speech and the morphological forms of the respective sense. This is followed by a sense definition and sense examples that illustrate both the use and the meaning of each particular sense. For the example sentences, make sure to use different sentence structures, referring to different people; refer to past, present, and future situations; vary long and short example sentences; and include other elaborations so that you, dear ChatGPT, produce high-quality dictionary material for me.

- Translation of Figure 3:

Thanks, but actually I wanted to see a monolingual dictionary entry for the Portuguese entry '*greve*' (which is 'strike' in English).

I apologize for the confusion earlier. Here is a monolingual dictionary entry for the word '*greve*' in Portuguese:

greve (feminine noun)

Frequency: The word *strike* is among the 10,000 most frequent words in Portuguese.

Definition: A *strike* is a stoppage organized by workers in which they stop working as a form of protest, usually to demand better working conditions, higher wages, or other rights from employers.

Examples:

1. Factory workers went on *strike* to demand a pay rise.
2. The teachers decided to go on *strike* because of the lack of investment in education.
3. The truckers' *strike* caused supply shortages across the country.

Note: The word 'strike' is often used in the context of labour protests and is a right guaranteed by labour laws in many countries. It can be carried out by various professional sectors, including teachers, doctors, drivers, and others, as a means of putting pressure on employers or governments to meet workers' demands.

- Translation of Figure 4:

Please give me a COBUILD-style monolingual dictionary entry for '*recomendar*'. Start with the frequency information at the top. Each sense should be in a numbered block. Each block then starts with the part of speech

and the morphological forms of the respective sense. This is followed by a sense definition and sense examples that illustrate both the use and the meaning of each particular sense. For the example sentences, make sure to use different sentence structures, referring to different people; refer to past, present, and future situations; vary long and short example sentences; and include other elaborations so that you, dear ChatGPT, produce high-quality dictionary material for me. Also include a pronunciation of the entry using IPA, and include a list of synonyms and antonyms at the senses where such lists are relevant.

Note that a final line was added to the prompt, as was done at the end of the initial study as well (De Schryver 2023b); highlighted in grey above.

- Translation of Figure 5a:

recomendar (verb)

Frequency: The word *recomendar* is among the 2,000 most frequent words in Portuguese.

Sense 1: *To express a favourable opinion about something or someone; to advise or suggest something to someone*

Morphological forms: recommend, recommends, recommending, recommended

Definition: When you recommend something, you express a positive opinion about it, often by advising or suggesting that someone else do it.

Examples:

1. I *recommend* this restaurant; the food is delicious.
2. She *recommended* a good book for me to read on vacation.
3. They are *recommending* that everyone wears masks to protect against the virus.

Pronunciation: /ʁe.ko.mẽ.'daʁ/

Synonyms (Sense 1): advise, suggest, indicate, endorse, approve

Antonyms (Sense 1): advise against, disapprove of, discourage

- Translation of Figure 5b:

Sense 2: *Presenting someone as fit or suitable for something, such as a job or position*

Morphological forms: recommend, recommends, recommending, recommended

Definition: In this sense, to recommend means to present someone as qualified for a job or position, indicating their skills, qualities or achievements.

Examples:

1. He was *recommended* for the position of manager because of his experience and leadership skills.
2. The professor *recommended* several excellent students for scholarships at the university.
3. The letter of recommendation *recommended* the candidate as highly competent for the job.

Pronunciation: /ʁe.ko.mẽ.'daʁ/

Synonyms (Sense 2): indicate, endorse, refer, endorse, sponsor

Antonyms (Sense 2): disapprove, disrecommend, criticize, censure

- Translation of Figure 5c:

Sense 3: *To prescribe or suggest something as a treatment or solution*

Morphological forms: recommend, recommends, recommending, recommended

Definition: In medical or professional terms, recommending involves prescribing or suggesting a treatment, course of action or solution for a specific condition or problem.

Examples:

1. The doctor *recommended* rest and medication to treat the flu.
2. The specialist *recommended* a balanced diet and exercise to improve the patient's health.
3. She *recommended* that he consult a lawyer to resolve the legal issue.

Pronunciation: /ʁe.ko.mẽ.'daʁ/

Synonyms (Sense 3): prescribe, advise, guide, suggest, indicate

Antonyms (Sense 3): advise against, disallow, disapprove of, discourage

- Translation of Figure 9:

Give me a bilingual Portuguese–English dictionary entry for the word 'banco'. Start with the frequency information at the top. Each sense should be in a numbered block. Each block then starts with the part of speech and the morphological forms of the respective sense. This is followed by a sense definition in Portuguese, and translation of that definition into English. Note that to define, you have to use the full-sentence defining style, which consists of two clauses: the first clause illustrates the usage pattern, the second clause is a paraphrase of meaning. Also list one or more translation equivalents in English at each sense. Add Portuguese examples with their English translations that illustrate both the use and the meaning of each particular Portuguese sense. For the example sentences, make sure to use different sentence structures, referring to different people; refer to past, present, and future situations; vary long and short example sentences; and include other elaborations so that you, dear ChatGPT, produce high-quality dictionary material for me. Also include a pronunciation of the entry using IPA, and include a list of synonyms and antonyms at the senses where such lists are relevant.

The grey highlight added on top of an extra sentence in the prompt here, marks the attempt to make sure a COBUILD-style definition, so a full-sentence definition (FSD), is obtained by ChatGPT-3.5.

- Translation of Figure 10a:

bank [ˈbɛ̃ku]

*Frequency: ******

1. Noun (Financial Institution)

- **Morphological forms:** banks (plural)
- **Sense:** A financial institution where people and companies can deposit money, obtain loans and carry out various financial transactions.
 - *Usage Pattern:* A bank is a place where people and companies can deposit money, obtain loans and carry out financial transactions.
 - *Definition:* A bank is an establishment where people and businesses can put their money, get loans and manage various financial transactions.
- **English Translation Equivalents:** Bank
 - *Example 1 in Portuguese:* I need to go to the bank to withdraw some money.
 - *Translation 1 in English:* I need to go to the bank to withdraw some money.
 - *Example 2 in Portuguese:* The bank offers various financial services, including savings accounts and mortgages.
 - *Translation 2 in English:* The bank offers various financial services, including savings accounts and mortgages.
- **Pronunciation:** [ˈbɛ̃ku]
- **Synonyms:** financial institution, bank branch
- **Antonyms:** customer, borrower

- Translation of Figure 10b:

2. Noun (Bank or Sloping Area)

- **Morphological forms:** banks (plural)
- **Sense:** The side of a river, lake or other body of water; often referring to the raised area or the margin.
 - *Usage Pattern:* The margin of a river is the side of the river.
 - *Definition:* The margin of a river, for example, denotes the land next to the river.
- **Translation Equivalents in English:** Bank, Riverside
 - *Example 1 in Portuguese:* They decided to have a picnic on the margin of the river.
 - *Translation 1 in English:* They decided to have a picnic on the bank of the river.
 - *Example 2 in Portuguese:* The boat is anchored near the margin of the lake.
 - *Translation 2 in English:* The boat is anchored near the bank of the lake.
- **Pronunciation:** [ˈbɛ̃ku]
- **Synonyms:** margin, edge
- **Antonyms:** centre, middle

- Translation of Figure 10c:

3. Verb

- **Morphological forms:** *banco (3rd person singular present tense), *bancou (past tense), *bancando (present participle), *banco (3rd person singular present tense)
- **Sense:** To deposit or store something in a bank, or to entrust or store something for future use.
 - *Usage Pattern:* Someone *banks something in a bank, or they *bank something for a specific purpose.
 - *Definition:* When someone deposits something in a bank, they are storing it in a financial institution or entrusting it for future use.
- **English Translation Equivalents:** Deposit, Rely on
 - *Example 1 in Portuguese:* She deposits her savings in a local bank.
 - *Translation 1 in English:* She banks her savings in a local bank.
 - *Example 2 in Portuguese:* He's counting on you winning the competition with your skills.
 - *Translation 2 in English:* He's banking on winning the competition with his skills.
- **Pronunciation:** ['bẽku]
- **Synonyms:** deposit, entrust
- **Antonyms:** withdraw, discredit

Note that the stars (*) were added to this translation, to mark (some of) the various errors in Portuguese. See the main text for the explanation.

http://lexikos.journals.ac.za; https://doi.org/10.5788/35-2-2108 (Volle uitgawe / Full volume)
https://lexikos.journals.ac.za; https://doi.org/10.5788/35-2-2103 (Article)

Addendum B: Online Swahili–English Dictionary / Kamusi ya Kiswahili–Kiingereza

- Without support, the following online Swahili–English dictionary — having been free and top-ranked by Google⁷ for over two decades — will have to be taken down (<https://africanlanguages.com/swahili/>).

The screenshot shows a web browser window with the URL <https://africanlanguages.com/swahili/>. A prominent red-bordered box contains a message: "Dear Swahili dictionary user, Unfortunately, we are out of funds for hosting costs for this site and urgently need donations to keep it running. Please consider donating at PayPal.me/davidjoffe. Thank you! If you would like to sponsor this site (for example, with a retail advertising banner here), [Contact Us](#)". Below this, there are links for "New 3 Oct 2023: New article on [Generative AI and Lexicography](#)", "Song Trein na Pretoria", and "Creative Gardening with indigenous Plants". A white box with a blue border contains a PayPal donation form with the text "Please donate now with Payfast to help keep this Swahili dictionary online and going - thank you. - David Joffe". The form has an "Amount (Rands) (R80 = 80)" field, a "Quantity: 1" field, and a "Pay Now" button. Below the form are two buttons: "PayPal donations" and "Bitcoin donations". A blue header reads "Online Swahili - English Dictionary" with a language selector for "Kiswahili | English". A search bar contains "search / tafuta" and a "Search" button. A welcome message says "Welcome to africanlanguages.com/swahili/ (now more mobile-friendly!) please give the mobile version a try". A link says "If you would like to advertise on this page, please [contact us](#)". A grey box contains a link: "Click here to try or buy a **downloadable Windows version** of this dictionary » (Sales help support this dictionary)". At the bottom, a news item states: "News 13 Oct 2014: This dictionary now also returns results from the Microsoft Swahili terminology list, below the regular results!"

Addendum C: Using ChatGPT-3.5 — For Swahili Lexicography [here the translations into English]

- Translation of Figure 15:

Give me a bilingual Swahili–English dictionary entry for the word '-pendekeza'. Start with the frequency information at the top. Each sense should be in a numbered block. Each block then starts with the part of speech and the morphological forms of the respective sense. This is followed by a sense definition in Swahili, and translation of that definition into English. Note that to define, you have to use the full-sentence defining style, which consists of two clauses: the first clause illustrates the usage pattern, the second clause is a paraphrase of meaning. Also list one or more translation equivalents in English at each sense. Add Swahili examples with their English translations that illustrate both the use and the meaning of each particular Swahili sense. For the example sentences, make sure to use different sentence structures, referring to different people; refer to past, present, and future situations; vary long and short example sentences; and include other elaborations so that you, dear ChatGPT, produce high-quality dictionary material for me. Also include a pronunciation of the entry using IPA, and include a list of synonyms and antonyms at the senses where such lists are relevant.

- Translation of Figure 16:

-pendekeza

Repeat: -pendekeza (Verb)

1. Verb - To do

Meaning: To give an opinion or recommendation to do something because you believe it is good or appropriate to do it.

Translation: recommend

Synonyms: to recommend, advise

Antonyms: to reject, to oppose

Examples:

- *The doctor recommended that the patient rest more.*
(The doctor recommended that the patient rest more.)
- *The teacher recommended that we do extra research for the upcoming exam.*
(The teacher recommended that we do extra research for the upcoming exam.)

Pronunciation: /pende'keza/

Synonyms and Antonyms:

- Synonyms: to recommend, advise
- Antonyms: to reject, oppose

Customised GPTs for Lexicography

Gilles-Maurice de Schryver, *BantUGent & LT³,
Ghent University, Ghent, Belgium;
and Department of African Languages,
University of Pretoria, Pretoria, South Africa*
(gillesmaurice.deschryver@UGent.be)
(<https://orcid.org/0000-0001-7272-9878>)

Abstract: About a year and a half ago, lexicographers realised that the enthusiasm they had when using ChatGPT-3.5 for English lexicography, could not be carried over to other languages, not even big ones like Portuguese. A year ago, the lexicographic community was made aware that ChatGPT-3.5 could not be used at all for small languages, and especially not for oral ones like Lusoga. In mid-2025, we notice that the models have improved drastically. Both Gemini-2.5 and especially GPT-4o perform miracles on languages like Portuguese, and for undocumented languages like Lusoga, customising GPT-4o by feeding it corpora as well as entire dictionaries is a major step forward, down to tackling neologisms. We are not yet there in full but moving forward.

Keywords: GENERATIVE ARTIFICIAL INTELLIGENCE (GENAI), LARGE LANGUAGE MODELS (LLMs), GENERATIVE PRE-TRAINED TRANSFORMER (GPT), CUSTOMISED GPTs, LEXICOGRAPHY, DICTIONARIES, LEXICOGRAPHERS, PORTUGUESE, LUSOGA

Obufunze: GPT emmumbe eya GenAI mw'isomo ly'okuwandiika amawanika.

Mu ibbanga nga ely'omwaka mulala n'ekitundu eliweile, abawandiisi b'amawanika baatandiika okuteebeleza nti eitutumu lye baatuukaku nga bakozeza ChatGPT-3.5 ku lulimi Olungeleza, tilisobola kwongezebwayo kukola ku nnimi dhindi ng'Olupootigo. Omwaka oguweile, abawandiisi b'amawanika bano beene b'amala b'ayolekebwa nti ChatGPT-3.5 tesobola kukozezebwa ku nnimi dhitamanhiibwa inho, nailala ng'edho edhikaasinga okuba obw'omu ndhogela ng'Olusoga. Ebyafaayo by'ensonga dhino dhombi waigulu byandhulwa mu kitundu ky'olupapula luno ekisooka. Kino n'ekitundu eky'okubili ekilaga ebibailewo okutuuka mu magatigati g'omwaka 2025, ekilaga nti enteeko ya tekinologiya akozesebwa mu ChatGPT-3.5, ayongeile okwiluulibwamu. Kino kiidhie oluvainhuma lw'okugezezebwa okulaze nti, olwaleelo luno, enkola dha Gemini-2.5 ni GPT-4o dhiyumbwilwa inho mu ngeli ye dhikola ku lulimi ng'Olupootigo okusingila ilala. Bwe kiila ku nnimi edhitali mu buwandiike ng'Olusoga, okuliikiliza GPT-4o ebiwandiiko n'ebigambo ebiva mu itu ly'Olusoga, awalala n'ebyo ebiva mu nkenga dh'Eiwanika ly'Olusoga, kusenvwilwa kinene inho kw'aba nti kusobozesa n'ennondoola y'ebigambo ebiyaaka ebyakaingila mu lulimi luno. Waile ng'enkola eyeefaanaanhiliza kw'eyo eliwo buti ku Lungeleza n'Olupootigo ekaali kutuukibwaku mu Lusoga, aye, olugendo olukiluubilila lutandiikiibwaku.

Ebigambo ebikulu: KALIMAGEZI-KISIBUKILA, OKUTETENKELEZA KW'ENNIMI ENNHINGI, KISIBUKILA AKAALI KUKENGULWA, GPT EMMUMBE, EISOMO LY'OKUWANDIIKA AMAWANIKA, AMAWANIKA, ABAWANDIISI B'AMAWANIKA, OLUPOOTIGO, OLUSOGA

1. Out-of-the-Box GPTs

1.1 Chatbot Arena

Even though Chatbot Arena¹ now exists where one can currently pair any two of about two-hundred LLMs — and thus choose from models like **Gemini** (Google), **GPT** (OpenAI), **Claude** (Anthropic), **DeepSeek** (High-Flyer), **Grok** (xAI), **Qwen** (Alibaba), **LeChat** (Mistral), **Hunyuán** (Tencent), **Command R+** (Cohere), **Gemma** (Google), **GLM** (Zhipu), **Step** (StepFun), **LLaMA** (Meta), **Yi** (01.AI), **Amazon Nova** (Amazon), etc. — OpenAI's GPT models still have the strongest conversational abilities, content creation, and summarisation, while Google's Gemini models are top for real-time web searches, complex reasoning, and research tasks. OpenAI's GPT and Google's Gemini models will therefore be used in this article.

1.2 Semi-visible Portuguese

The availability of the Chatbot Arena thus means that any of the prompts designed in earlier lexicographic research may now be re-used to test the quality of the output of over 200 models, to see which one works best for which type of dictionary, which language, etc. While this clearly provides for near-endless fun, all of these, as offered, are still out-of-the-box (see De Schryver 2025). Apart from the obvious improvement to have easy access to so much variety, one should check whether some of the recent LLMs are still as bad for non-English as those from a year and a half ago, as was the case for Portuguese when the then ChatGPT-3.5 was used at the end of October 2023 (De Schryver 2023b).

As examples, Figures 1 to 4 show a comparison between the Gemini-2.5 model and the GPT-4o model, for the prompts shown at the top-right in each case, to get to proper dictionary articles for: (1) the verb of communication *recommendar* in an English–Portuguese learners' dictionary, (2) the lemma *table* in an English–Portuguese learners' dictionary, (3) the lemma *table* in a Portuguese–English learners' dictionary, with the prompt in English, and (4) the lemma *table* in a Portuguese–English learners' dictionary, with the prompt in Portuguese. The full texts offered by both the Gemini and GPT models, received as-is and thus left untouched, are presented in Addendum A, available online, as Figures 1 to 4 only show the opening sections. Click [here](#) for Addendum A.

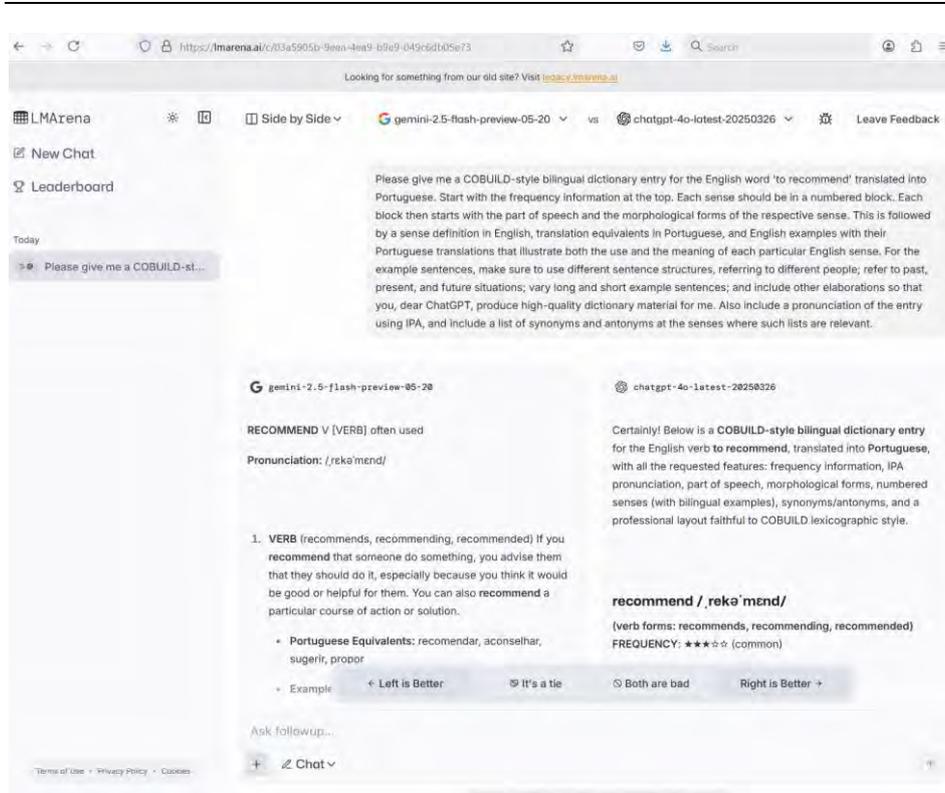


Figure 1: Using Chatbot Arena to compare any two models for lexicographic purposes, here Google's *gemini-2.5-flash-preview-05-20* vs. OpenAI's *chatgpt-4o-latest-20250326*, for the verb of communication *recommend* in an English–Portuguese learners' dictionary

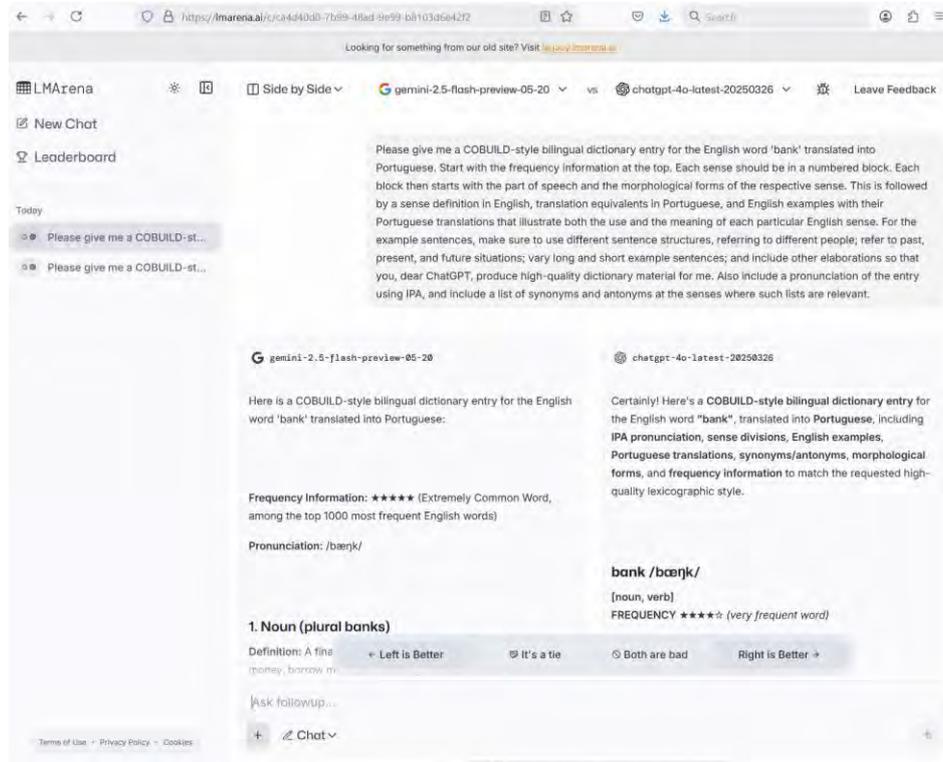


Figure 2: Using Chatbot Arena to compare any two models for lexicographic purposes, here Google's *gemini-2.5-flash-preview-05-20* vs. OpenAI's *chatgpt-4o-latest-20250326*, for the lemma *table* in an English–Portuguese learners' dictionary

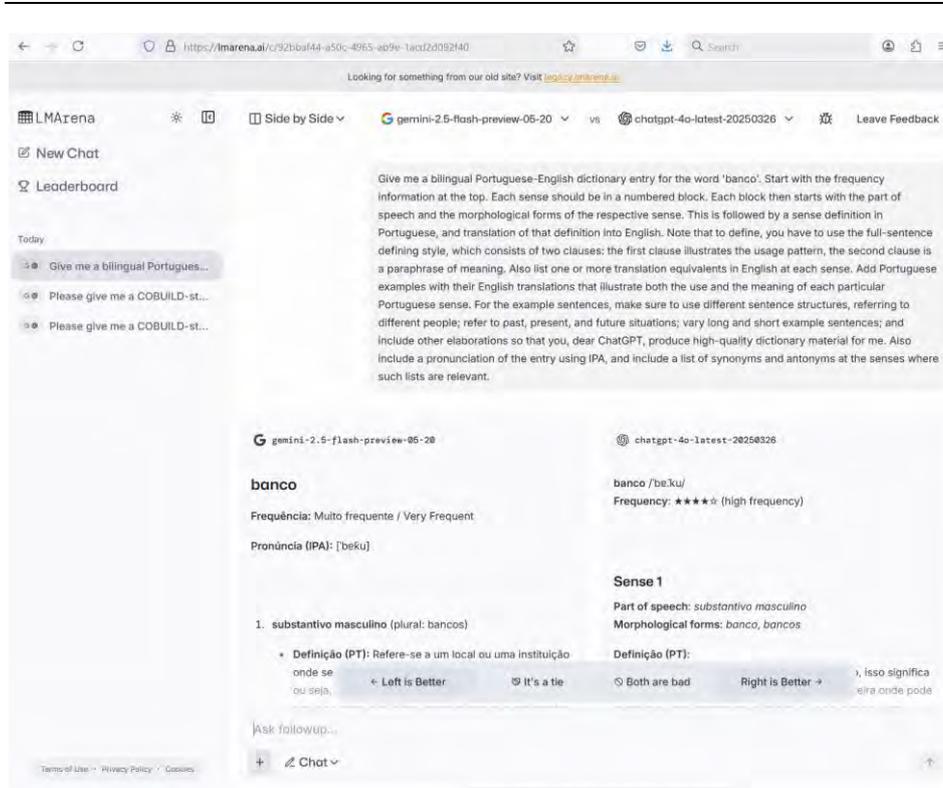


Figure 3: Using Chatbot Arena to compare any two models for lexicographic purposes, here Google's *gemini-2.5-flash-preview-05-20* vs. OpenAI's *chatgpt-4o-latest-20250326*, for the lemma *table* in a Portuguese-English learners' dictionary, with the prompt in English

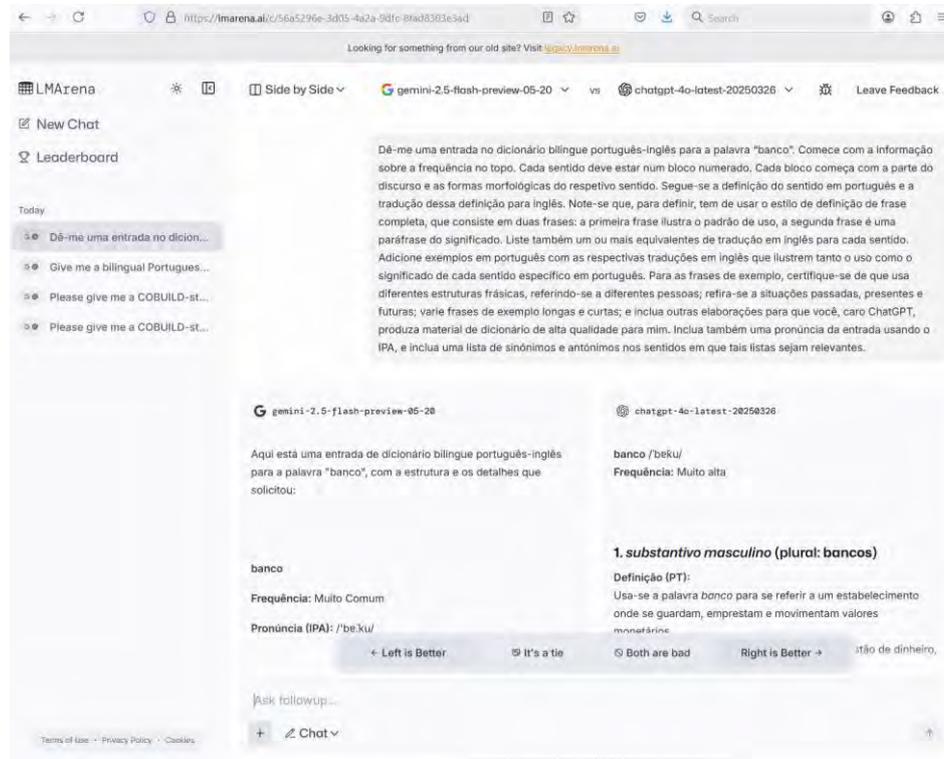


Figure 4: Using Chatbot Arena to compare any two models for lexicographic purposes, here Google's *gemini-2.5-flash-preview-05-20* vs. OpenAI's *chatgpt-4o-latest-20250326*, for the lemma *table* in a Portuguese–English learners' dictionary, with the prompt in Portuguese

Astonishingly, all the problems ChatGPT-3.5 had when trying to treat non-English like Portuguese for lexicographic purposes (see De Schryver 2025) are simply gone. This is most evident from the treatment of the lemma *table* when prompting GPT-4o, no matter the direction of the dictionary, no matter whether the prompt is in English or in Portuguese. As of June 2025 no more hallucinations are evident, and only noun senses are given, and thus no more inventions of non-existent verbal uses of Portuguese *banco*. Given that there are four prompts, submitted to two LLMs, there are a total of 8 answers in Addendum A, available online, numbered A1 to A8. Click [here](#) for Addendum A. More interesting summaries now follow for each of the eight dictionary articles produced via GenAI.

In A1, Gemini-2.5 for *recommend*, we immediately note excellent COBUILD-style full-sentence definitions (FSDs). On top, at sense 3, following the FSD, a second sentence starts with 'This usage often implies that ...' — brilliant, as indeed, meaning is the result of usage, with meaning potentials only triggered by the

context, as Patrick Hanks reminded us throughout his life (Hanks 1979, 1988, 2000, 2002, 2012, Hanks and Franklin 2019, Hanks and Može 2019, Grefenstette and Hanks 2023). The Portuguese translations of the examples at sense 3 are also brilliant, as the various translation equivalents offered are indeed used (the underlinings, both full and broken up, were added by the present author). Lastly, the separation into three uses, and thus three senses, makes perfect sense.

In **A2**, GPT-4o for *recommend*, and thus even though the same verb is treated, the three uses/senses are not the same as in A1. In a way, the uses here are even further apart, as is also obvious from the additions of [VERB + that] at sense 2, and [VERB + itself, be + VERB] at sense 3. All of this implicitly refers to Patrick Hanks's Corpus Pattern Analysis (Hanks 2004, Hanks and Pustejovsky 2005, Hanks 2008, Hanks et al. 2018, Hanks and Ma 2021), and more specially his (unfinished) *Pattern Dictionary of English Verbs*². As impressive are the Usage Notes, which fully delve into linguistics, and the Related entries linking the verb to the noun and adjective.

In **A3**, Gemini-2.5 for *bank*, the first sense (as a financial institution) and second sense (as the sloping land alongside a river) mention that there are 'no direct antonyms' with a reason why — smart. At the second sense the mistake is not made to include **banco* as one of the translation equivalents, and as to the synonym 'levee' the context is added between brackets: '*for a raised bank to prevent flooding*' — very good. At the third sense (as a mass or pile of something) the translation equivalents are provided with context — excellent. At the fourth sense (as the verb *to bank*) the definition is sadly not COBUILD-style, but the mistake is not made to include **banco* as a translation equivalent. The fifth sense (of an aircraft or vehicle) is even far-more detailed than as seen in the actual COBUILD.³ The sixth sense (a row or series of similar items) correctly illustrates two of the translation equivalents. At the seventh sense, we see all sorts of correct compounds. At the eighth sense the phrasal verb *bank on* is treated, but alas no FSD. The examples here use two different translation equivalents.

In **A4**, GPT-4o for *bank*, the first sense (as a financial institution) *does* include antonyms — even smarter than Gemini-2.5. At sense two (as the side or edge of a river) only *margem* is (correctly) offered as a translation equivalent. Sadly, at the third sense (as the verb *to bank*), no proper FSD, and all examples use only one of two possible translation equivalents. At the fourth sense (a row or series of similar things) the third example uses a translation equivalent not listed; the other two (correctly) use two different ones. All is well with the fifth sense (the noun in aviation or driving). Idem for the sixth sense (the verb in aviation or driving), and seventh sense (as a mass or layer of something). The 'Summary Usage Note' is most-excellent and reminds the user of the highly polysemous nature of this lemma, as it is used both literally and metaphorically, across various fields, pointing out that the resulting translation equivalents in Portuguese truly depend on context. Brilliant lexicography.

In **A5**, Gemini-2.5 for Portuguese *banco* (prompt in English), at the first sense (as a financial institution) the translation equivalent and use in example sen-

tences are all correct. No antonym is provided, although we know, thanks to GPT-4o under A4 sense 1 that there are antonyms (namely *borrower*, *debtor*). At the second sense (as furniture for sitting), three translation equivalents are provided and two are used in the examples (*bench*, *seat*). At the third sense (as a sandbank, or school of fish), four translation equivalents are provided and all four are used in the examples (*sandbank*, *school*, *bank*, *shoal*) — good lexicography in a learners' dictionary. At the fourth sense (as a blood or organ bank), only bank is offered as a translation equivalent, but all examples combine it (once each with *blood*, *organ*, *gene*, *tissue*). None of senses 2 to 4 include antonyms; but GPT-4o does for all, see the next example.

In A6, GPT-4o for Portuguese *banco* (prompt in English), the definition at the first sense (as a financial institution) is a good attempt at an FSD, which is always tougher for a noun than for a verb. Two translation equivalents are offered, but only one is used in the three examples (*bank*). As antonyms of this use of *banco*, Portuguese *cofre* 'safe' and *colchão (em uso figurado para guardar dinheiro em casa)* 'mattress (figuratively for storing money at home)' are offered — very good. At the second sense (as a bench in a park), the FSD is near-excellent. Two translation equivalents are offered, of which only one is used in all three examples (*bench*). Here too, the antonyms are head-on, Portuguese *chão* 'floor' and *cadeira individual* 'single chair'. At the third sense (as a substitute bench), the FSD is again near-excellent. While two translation equivalents are offered, again only one is used in all three examples (*bench*), though here it makes more sense, as the expanded *substitute bench* is of course rarer. The antonyms are again very good, Portuguese *titular* 'starter' and *escalado* 'line-up'. At the fourth sense (as a database), we see the usual near-FSD, two translation equivalents of which one used in the three examples (*database*), while the antonyms are original, Portuguese *ausência de dados* 'no data' and *sistema manual* 'manual system'. At the fifth sense (as a sandbank), again a near-FSD, three translation equivalents, but in the three examples only one of them (twice *sandbank*, plus even *bank*, which is not among the translation equivalents), and as antonyms the debatable Portuguese *canal* 'canal' and *profundidade* 'depth'. The concluding summary at the end, focusing on the polysemy of Portuguese *banco*, found with financial, physical, sportive, digital, and geological uses — listed in the order of the five senses — is most precise and helpful. That note continues with the claim that this GenAI entry is 'offering a comprehensive bilingual reference for language learners and lexicographers alike'. Compared with the ridiculously short entry for *banco* as offered by Collins,⁴ this is undoubtedly better, near-perfect even. Near-perfect, as the FSD does indeed contain the two parts, but is followed by a semi-colon and then a more traditional definition on top, making the FSD a bit too long. Other than that, let GPT-4o compile an entire dictionary in this format overnight, for the full lexicon, and we might have (the draft of) a good bilingual Portuguese–English learners' dictionary right away.

In A7, Gemini-2.5 for Portuguese *banco* (prompt in Portuguese), even more is presented in Portuguese as a result of prompting in Portuguese. At the first

sense (as a financial institution), the definition consists of two sentences, so not an FSD, but if the two had been merged and shortened, it comes close. Of the two translation equivalents, all four examples use one (*bank*). Two synonyms are offered, no antonyms. At the second sense (as a bench in a park), we see the same type of non-FSD. Of the three translation equivalents that are offered, two are used in the four examples (*bench, pew*). Three synonyms are offered, no antonyms. At the third sense (as a sandbank), again a non-FSD. Of the four translation equivalents that are offered, three are used in the three examples (*sandbank, mudflat, shoal*). Three synonyms are offered, no antonyms. At the fourth sense (as a blood or data bank), another non-FSD. Of the four translation equivalents that are offered, only one is used in the four examples (*bank*). Four synonyms are offered, no antonyms. At the fifth sense (as a school or shoal of fish), the same issue with the non-FSD is evident. Of the two translation equivalents that are offered, only two are used in the three examples (*school, shoal*). Only one synonym is offered, no antonyms.

In A8, GPT-4o for Portuguese *banco* (prompt in Portuguese), there is again more in Portuguese, so here too the result of the Portuguese prompt. The first sense (as a financial institution) presents a definition in the FSD-style, use vs. meaning, followed by a second sentence that is a variation on the meaning only. Surely, the two sentences could have been merged, to achieve the FSD-goal of a learners' dictionary. One translation equivalent is offered and used in all three examples (*bank*). Two synonyms follow, with this time an indication that there are no antonyms in this sense (even though GPT-4o offered some for this sense in A4 and A6). The second sense (as a long seat) presents a similar FSD cum extra sentence. Two translation equivalents are offered, but only one is used in all three examples (*bench*). Three synonyms follow, and two antonyms (*cadeira* 'chair' and *poltrona* 'armchair'). The third sense (as a subaquatic formation) again has the pseudo-FSD and then a second sentence on a reformulation of the meaning. Four translation equivalents are offered, of which two are used in two examples (*sandbank, bank*) and then one not listed under the translation equivalents (*school*). Three synonyms follow, and two antonyms (*mar profundo* 'deep sea' and *leito rochoso* 'bedrock'). The fourth sense (as a vehicle seat) once again has the pseudo-FSD and second sentence. Two translation equivalents are offered, and both are used in all three examples (*car seat, seat*). Two synonyms follow, and two antonyms (*espaço livre* 'free space' and *chão do veículo* 'vehicle floor'). The fifth sense (as a collection of information) repeats the pseudo-FSD plus extra sentence on meaning. Two translation equivalents are offered, and both are used in all three examples (*database, data bank*). Two synonyms follow, and two antonyms (*informação dispersa* 'scattered information' and *fonte não estruturada* 'unstructured source'). At the very end, a pronunciation is offered for the lemma (both in European and Brazilian Portuguese, although the same here), followed by an invitation to treat GPT-4o as a professional lexicographer, as it invites the user with: 'If you'd like examples of compositions with '*banco*' (such as *data bank, square bank, image bank*), I can draw up more specific blocks. Would you like to continue?'

Even a cursory look — but especially the more detailed study just presented — of the lexicographic output of Gemini-2.5 vs. GPT-4o, makes clear that for lexicography GPT-4o is the better tool. Comparing the output under A6 (with the prompt in English) vs. the output under A8 (with the prompt in Portuguese) further indicates that prompting in English remains better than prompting in another language. It is therefore not surprising that over the past two years and a half, lexicographers not only started with but also mainly stuck by ChatGPT, as seen from the detailed study of a sample of 100 contributions on 'GenAI in lexicography' analysed in De Schryver (forthcoming-a, forthcoming-b). In any case, for large languages like Portuguese there is thus no need anymore to customise an LLM, at least not for general lexicographic purposes.

2. Customised GPTs

The necessary move for large non-English languages — so from a 'hyper-visible' to a 'semi-visible' uptake in technology like GenAI — has been made: Recent LLM models are now also achieving top lexicographic performances for languages such as Portuguese with out-of-the-box tools. Of course these same tools won't have improved much for low-visible languages such as the Bantu languages, and very likely not for a local-only visible language such as Lusoga: As far as we know, no extra data was added anywhere online in recent years for that language, so for exotic oral languages like Lusoga it becomes important to now know whether a tool like GPT-4o can be used to take the next step for them, namely *customisation*.

2.1 Local-visible Lusoga

Since the release of GPT-4o on 13 May 2024, customisation has been put into the hands of users. Within days of the release of GPT-4o a serious attempt was made to customise this model by trying the impossible: Could it compile dictionary articles for *neologisms* in the oral — so extremely under-documented — Bantu language Lusoga, by feeding GPT-4o everything that exists on Lusoga, viz. a corpus of 3.7 million tokens (De Schryver and Nabirye 2022), and a full comprehensive monolingual dictionary in XML, the *e-Eiwanika ly'Olusoga* (Nabirye et al. 2012)⁵? The outcome, presented on 20 May 2024 as a keynote at the CogALex workshop of LREC-COLING 2024 in Turin (De Schryver 2024a), was tantalising: whereas GPT-4o generated only nonsense out-of-the-box, it started to make sense with this customisation.

Indeed, following an opening prompt in which GPT-4o was merely informed that the word *ente* is 'cow' in Lusoga; an out-of-the-box generation of a definition for it produced utter gibberish: the reply seen in Figure 5 is meaningless.



Figure 5: Trying the out-of-the-box GPT-4o, asking for a Lusoga definition of *ente* 'cow'

A team of three therefore joined forces — programmer David Joffe from South Africa, native Lusoga speaker Minah Nabirye from Uganda, and lexicographer Gilles-Maurice de Schryver from Belgium — and messaged the built-in GPT builder to create a custom GPT, which they called 'Lusoga Linguist'. After agreeing on a profile picture, the refinement of the context could start, with Figure 6 showing one of the goals.

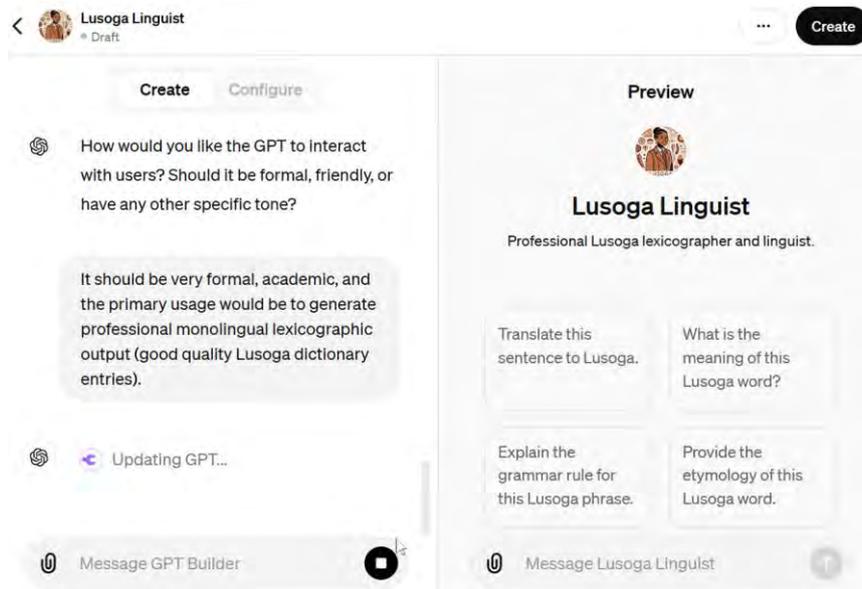


Figure 6: Work on refining the goal of the custom GPT for Lusoga lexicography has begun

After that, a first attempt was made at adding Lusoga corpus files to the 'Knowledge base', as seen in Figure 7.

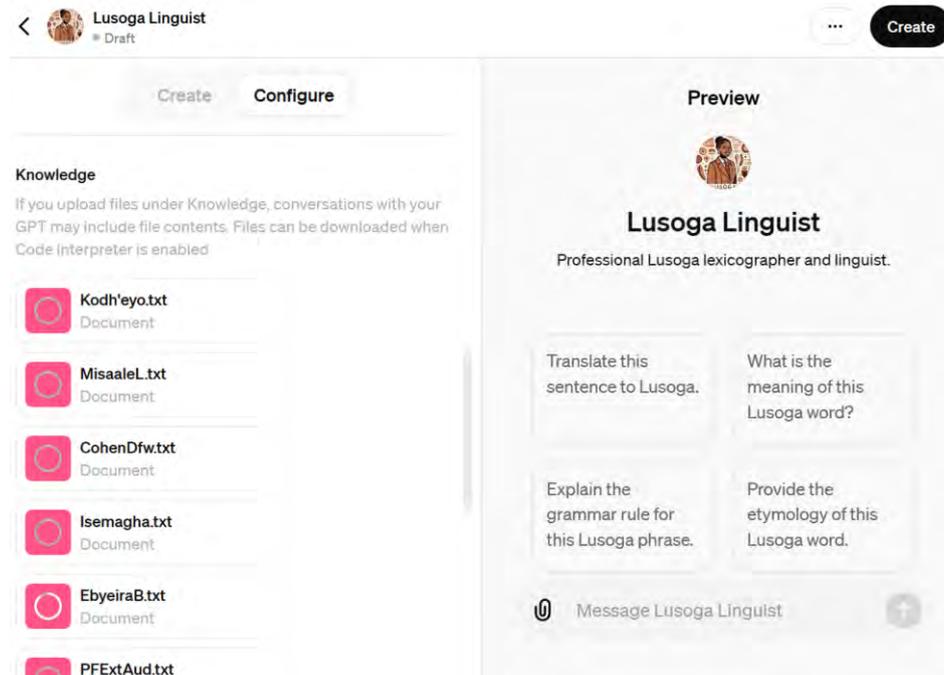


Figure 7: Adding a Lusoga sample corpus to the 'Knowledge base' of GPT-4o

Adding very many files did not work, so a single word file (of about 500KB) with 10,000 lines of corpus texts was added. In there the word *ente* 'cow' occurred 36 times. Once done, GPT-4o was instructed to '*study the uploaded corpus of Lusoga text*' to then use this knowledge before moving to the '*task of generating professional monolingual Lusoga dictionary entries*', as seen in Figure 8.

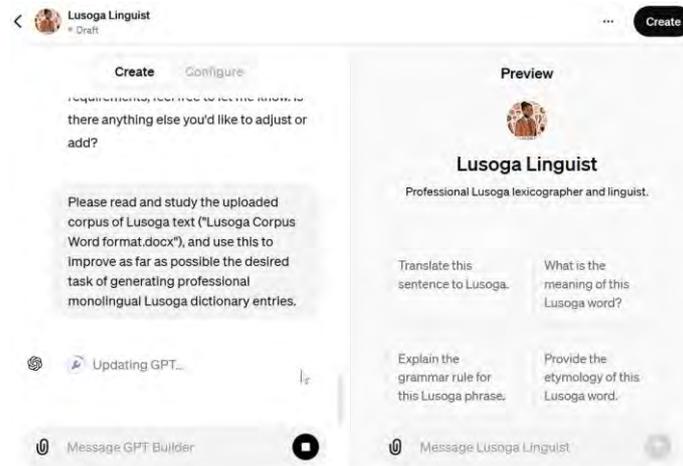


Figure 8: Instructing GPT-4o to study the submitted sample Lusoga corpus

The initial prompt was then repeated, as seen in Figure 9, together with the output.



Figure 9: A customised GPT-4o tries to define Lusoga *ente* 'cow', using a sample corpus

The output seen in Figure 9 reads more like a story than a dictionary. Plus, most is in the neighbouring and much larger language Luganda, with little in Lusoga. The opening of the definition translates as 'A cow is an animal that drinks milk and eats grass ...' and the last line of the screenshot has an example that translates as 'The cow of Waiswa gives birth to a chicken'. Further prompting in which GPT-4o was asked to only use Lusoga (and no Luganda) and to compile a proper dictionary (rather than a story) failed to improve the result. The decision was therefore taken to load samples of the monolingual Lusoga dictionary into the 'Knowledge base' of GPT-4o. That work was published as a paper dictionary (Nabirye 2009) as well as an e-dictionary (Nabirye et al. 2012), and for the latter both as a downloadable dictionary app as well as a free online dictionary. See Addendum B, available online (click [here](#)).

After uploading the sample with monolingual Lusoga dictionary examples, GPT-4o was instructed to study those sample entries, in the hope that it would 'understand' the lexicographic composition. Figure 10 displays the start of the conversation on this level, while Figure 11 displays the end of the analysis by GPT-4o. Recall that all the data submitted for training of the model is fully in Lusoga, not in any other language, so the task is clearly hard. On the other hand, there is a good lexical overlap between Lusoga and Luganda, a language for which there are considerable amounts of data online.

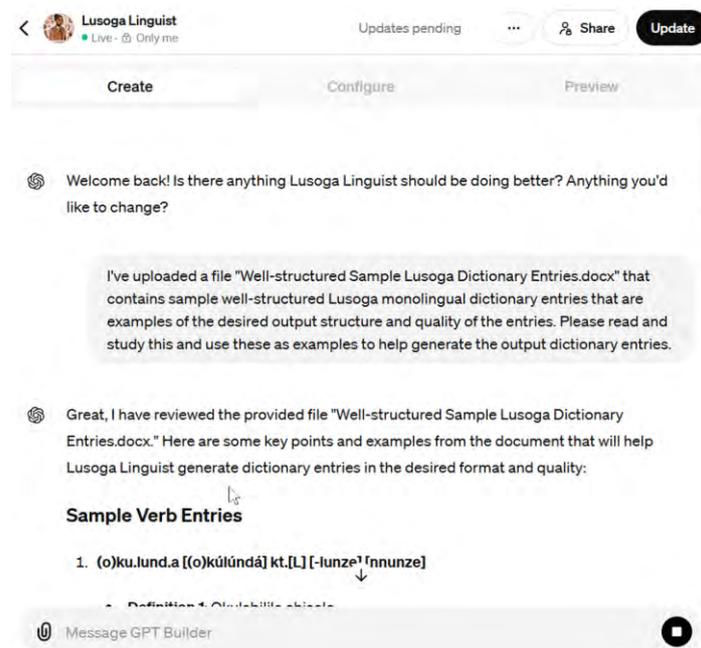


Figure 10: After adding a sample of monolingual Lusoga dictionary entries to the 'Knowledge base', asking GPT-4o to study the structure of that

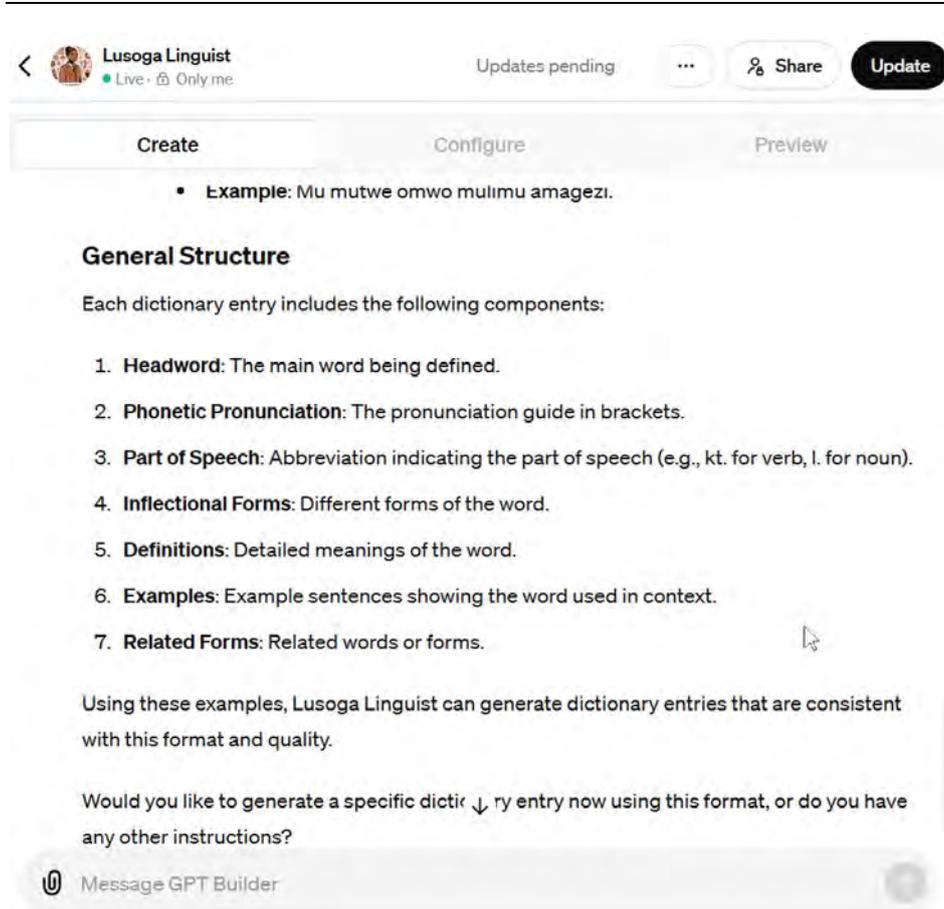


Figure 11: GPT-4o has 'learned' the lexicographic composition of entries in the monolingual Lusoga dictionary

Equipped with a sample Lusoga corpus, and a sample of Lusoga dictionary entries, GPT-4o was asked for a definition of *omukono* 'hand/arm' (not being among the sample dictionary entries), as seen in Figure 12. The overall appearance of Figure 12 is only slightly more dictionary-like. Plus, one continues to notice both Luganda and Lusoga, but whenever it is Lusoga it is still not coherent. Actually, but still surprisingly, only sense 2 is (1) written in Lusoga, and (2) makes somewhat sense, as the translation is 'One of the *drafts of the body ...'. Is the custom GPT-4o starting to learn Lusoga, or simply lucky in making things up? Two efforts were undertaken to make the custom GPT-4o much better.

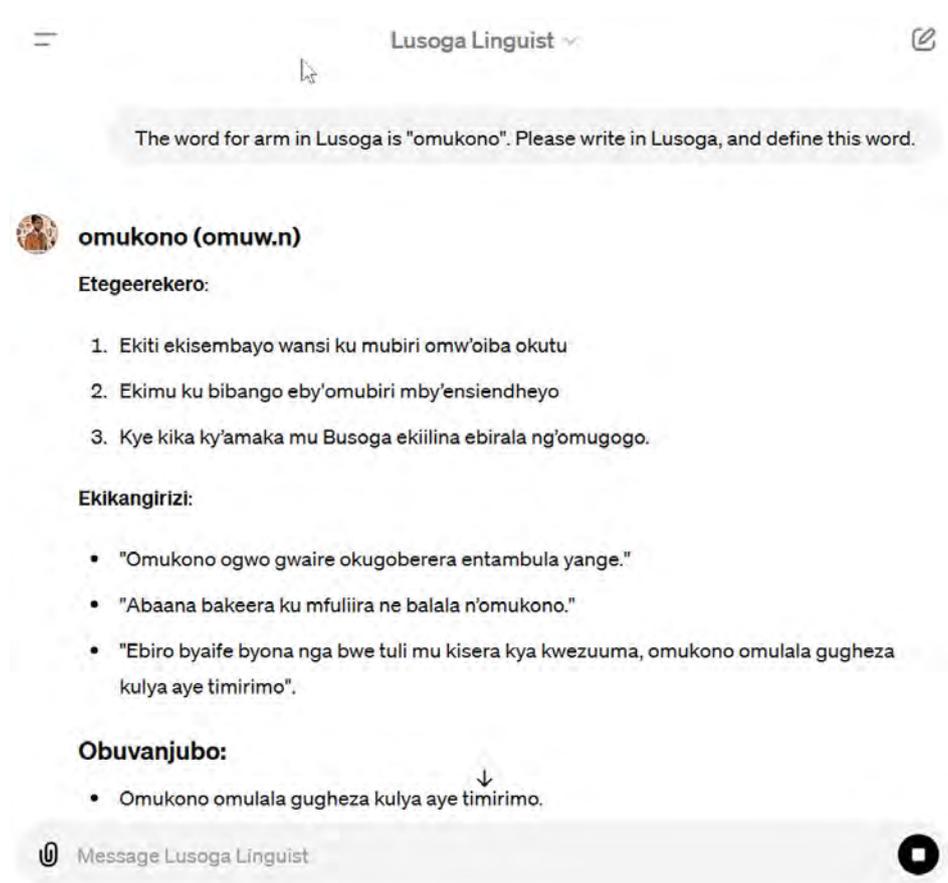


Figure 12: Asking the first version of a customised GPT-4o for the Lusoga entry *omukono* 'hand/arm'

First, given that up to 20 documents may be added to the 'Knowledge base', with each not containing more than 2 million LLM tokens and not larger than 512MB, all the files of the sample corpus were taken away and were now replaced with the full Lusoga corpus of 3.7 million word tokens, cut up in ten parts.⁶ That corpus was the result of about two decades of corpus building by Minah Nabirye (De Schryver and Nabirye 2018, 2022), and enlarging it even further is a truly heroic effort.⁷ Figure 13 shows how the 3.7 million Lusoga word tokens were added to the 'Knowledge base', and Figure 14 prompts with the same question as in Figure 12, but now assuming that the full corpus 'taught' the customised GPT-4o model more about the Lusoga language.

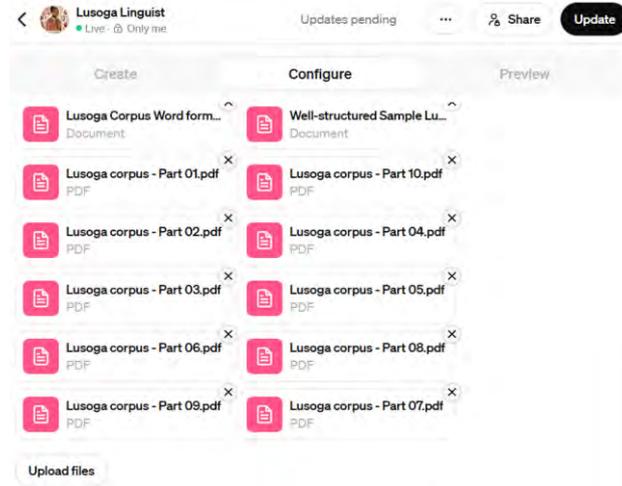


Figure 13: Adding all 3.7m word tokens of Lusoga to the 'Knowledge base' of the customised GPT-4o

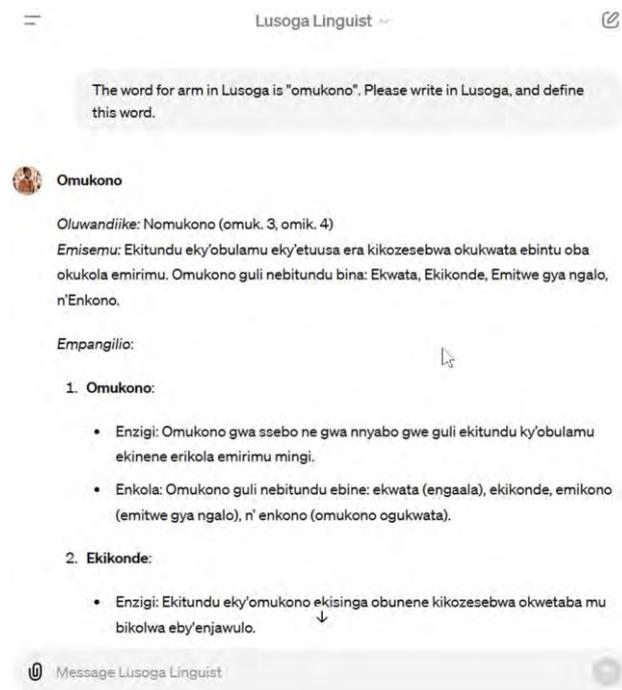


Figure 14: Asking the re-customised GPT-4o for the Lusoga entry *omukono* 'hand/arm'

While still far from correct, Figure 14 suddenly comes close to resembling real Lusoga and the analysis of Lusoga. On the first line, for instance, while the singular prefix is (obviously) correct (*omu-* in class 3), for the plural *omi-* in class 4 is suggested, where it should be *emi-* in class 4. Then, the definition of *omukono* 'hand/arm' contains not only too much Luganda, but also Swahili, and translates as '*The part of the body which is close and is used to hold things and to do work*'. A rather 'funny' definition, but it is valid: it is a close body part, and hands are used to hold things and also perform work. Compare this to the actual entry in the *e-Eiwanika ly'Olusoga*, as seen in the first screenshot of Addendum B, available online (click [here](#)).

This result logically led to the second effort, namely the addition to the 'Knowledge base' of the XML-file of the entire 600-page *e-Eiwanika ly'Olusoga*, in other words the full monolingual Lusoga dictionary, as seen in Figure 15. (The 'Well-structured Sample' was taken out.)



Figure 15: Adding the XML-file of the full monolingual *e-Eiwanika ly'Olusoga* to the 'Knowledge base' of the customised GPT-4o

This was immediately followed by an instruction to the custom GPT-4o to study this full dictionary, as seen in Figure 16.

🌀 Welcome back! Is there anything Lusoga Linguist should be doing better? Anything you'd like to change?

I have uploaded an XML file "Structured Eiwanika ly'Olusoga.xml" that contains a comprehensive monolingual Lusoga dictionary that is exactly an example of both the desired structure, fields, and type of entries that we want this GPT to produce. Can you read and study it, and read the entries, and use this all in order to generate dictionary entries, and then try to generate all dictionaries in future in this XML structure?

🌀 🔍 Searching my knowledge

Figure 16: Adding the XML-file of the entire monolingual *e-Eiwanika ly'Olusoga* to the 'Knowledge base' of the customised GPT-4o

Working with ChatGPT, Barrett (2023: 10) had noticed that even their GPT-4 model performed some stunts at times, namely to blatantly copy over existing dictionary contents wholesale (in his case for definitions of *virus*). In order to be sure that the customised GPT-4o model does not simply go on a copy-and-paste venture now that the full *e-Eiwanika ly'Olusoga* is in the 'Knowledge base', an additional difficulty was added, namely to focus on neologisms. The lemma *omuyumo* 'entertainer' was chosen, as it is a known neologism, first described in the academic literature in De Schryver and Nabirye (2022: 843), and *not* in the monolingual Lusoga dictionary, but *in* the Lusoga corpus. In Figure 17 one sees the reply to the question to compile a monolingual dictionary article for *omuyumo* 'entertainer'.

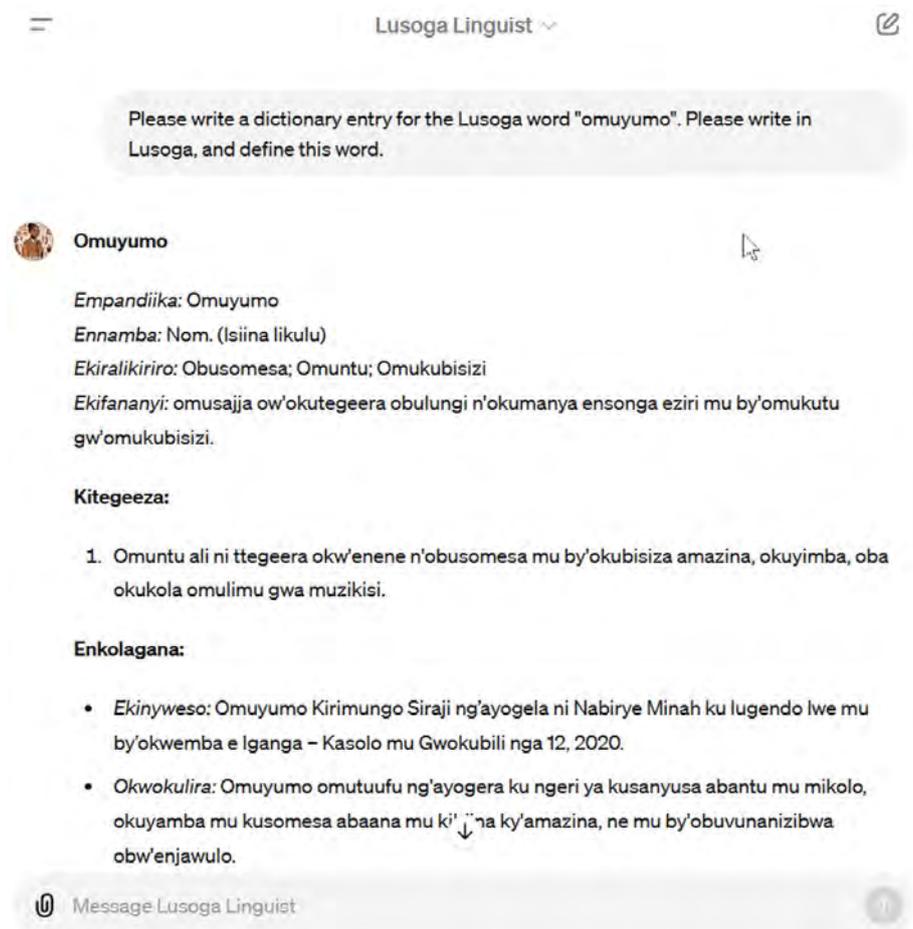


Figure 17: Asking the re-re-customised GPT-4o to compile a dictionary entry for Lusoga *omuyumo* 'entertainer'

While the opening section of Figure 17 still makes no lexicographic sense, the first bullet at the bottom provides an example that is fully grammatical, coherent and on-topic Lusoga. The translation of this example sentence is: '*The entertainer known as Kirimungo Siraj is in the process of speaking with Nabirye Minah about his Singing Journey at Iganga, Kasolo on 12-02-2020*'. This sentence does not appear in the corpus, and is clearly an excellent summary of one of the transcribed recordings made in 2020 for a book that came out four years later, ironically called *Obuyumu mu Busoga 'Entertainment in Busoga'* (Nabirye 2024). Unfortunately, while the example at the next bullet is again grammatical, coherent, and on-topic, it is couched in Luganda: '*A true entertainer who is speaking about ways of entertaining people at a function, who helps to teach the truth to children in class and for different types of responsibilities*'.

While there is clearly still a long way to go to get to proper lexicography, and while the neighbouring language Luganda still creeps in, seeing that GPT-4o can actually start learning a brand-new language and generate some dictionary material in it is promising, and, as far as we know, has not yet been tried for any other language within the field of lexicography.⁸

Moreover, when asked to explain where the data for the generated examples come from, GPT-4o is able to point to the various corpus sections from where it sourced and summarised information. For instance, for the example in Luganda at the last bullet in Figure 17, Figure 18 adds the Lusoga sources — in other words, here the example is based on material taken from the Lusoga corpus, but strangely 'translated' from Lusoga into Luganda.

- Ekinyweso: Omuyumo Kirimungo Siraji ng'ayogela ni Nabirye Minah ku lugendo lwe mu by'okwemba e Iganga – Kasolo mu Gwokubili nga 12, 2020.

- Okwokulira: Omuyumo omutuufu ng'ayogera ku ngeri ya kusanyusa abantu mu mikolo, okuyamba mu kusomesa abaana mu kibiina ky'amazina, ne mu by'obuvunanizibwa obw'enjawulo 【21:0†source】
【21:1†source】 【21:2†source】 【21:3†source】 【21:5†source】
【21:6†source】 【21:7†source】 【21:10†source】 .

Figure 18: The custom GPT-4o reveals the Lusoga sources used for the last example in Figure 17

Experimenting more with trying to discover what was used to come up with certain definitions and/or examples, at some point later on, while working on the neologism *obukurisitu* 'Christianity', the custom GPT-4o claims all of what is seen in Figure 19, where the most important claim is: '*Each reference helps trace*

back the information used in the dictionary entry to its original location in the uploaded documents. This ensures that the definition provided is well-supported by the source material'. And a bit later yet again, when working on the neologism *omukwampa* 'the direction of', the custom GPT-4o even goes as deep as to claim all that is seen in Figure 20, where the summary is: 'The references are not from a single contiguous line or section but rather from different parts of the text that discuss related topics, such as boundary demarcations, geographical descriptions, and historical or cultural contexts'. All of this has indeed a lot to do with the meaning of the neologism *omukwampa* 'the direction of', so if all true, this means that a customised LLM stops being the full black box feared by lexicographers, as it becomes somewhat less opaque. At the same time, customisation for an unknown exotic language such as Lusoga is not good enough, so we need more.

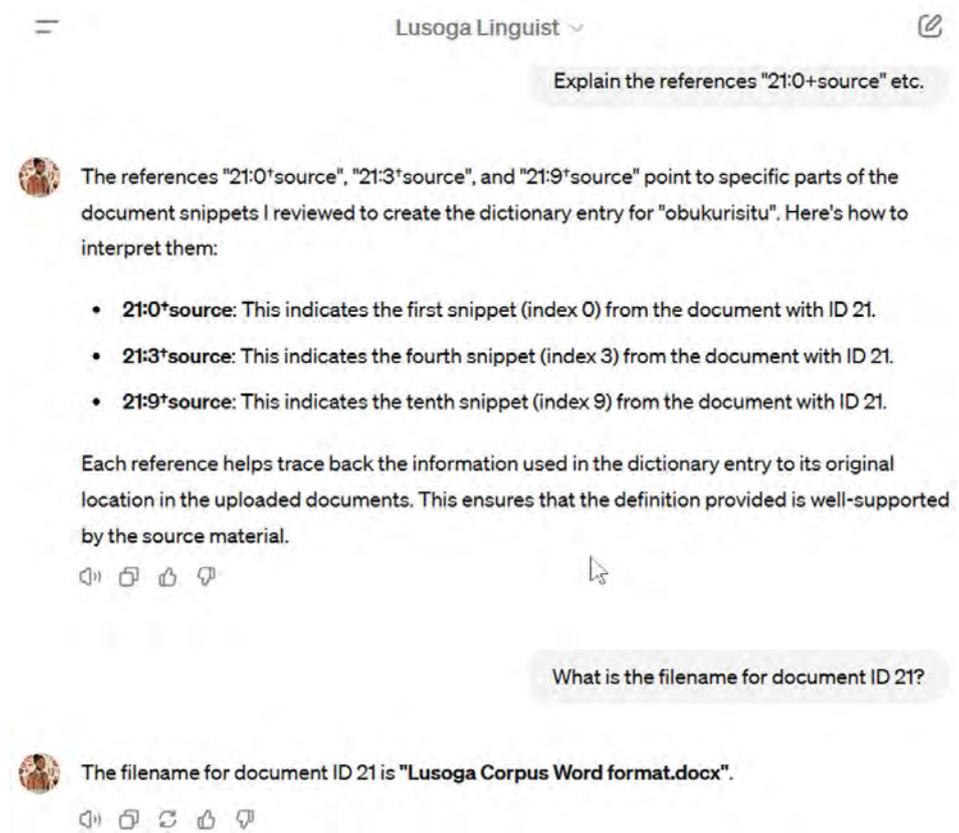


Figure 19: The custom GPT-4o reveals the Lusoga sources used for the Lusoga lemma *obukurisitu* 'Christianity'

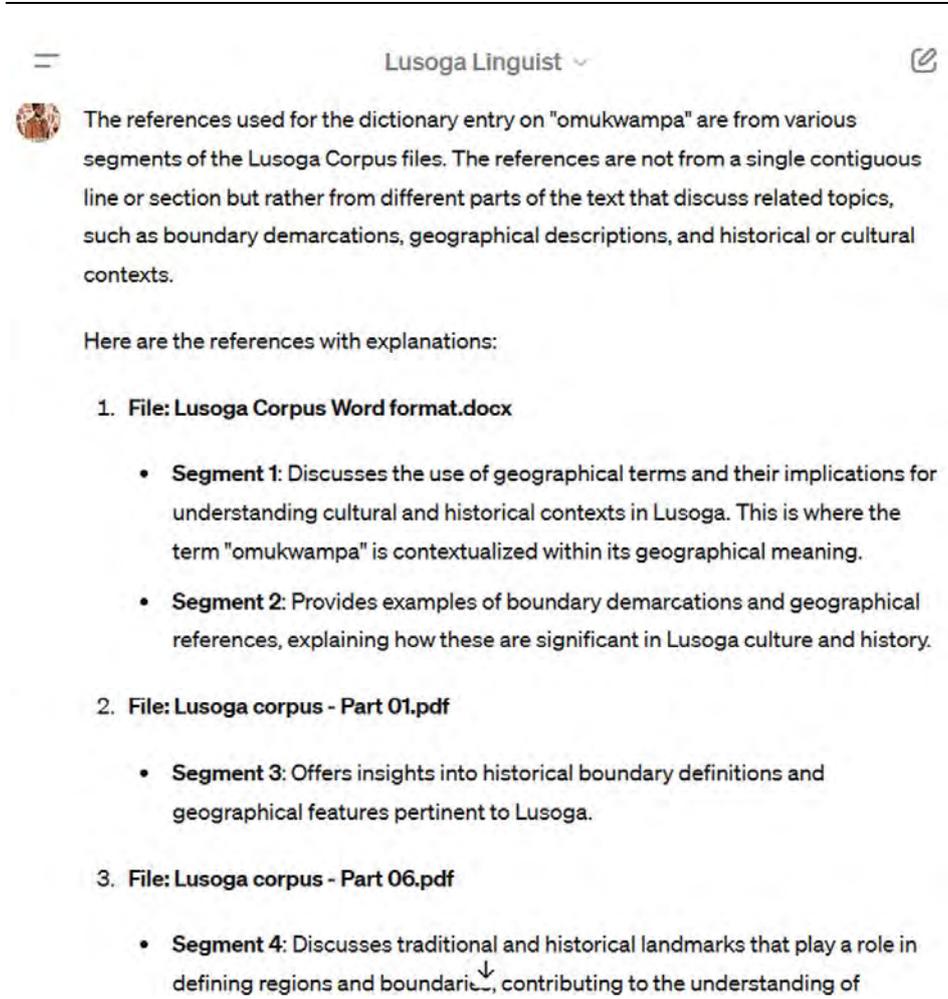


Figure 20: The custom GPT-4o reveals the Lusoga sources used for the Lusoga lemma *omukwampa* 'the direction of'

3. Epilogue

Armed with all the facts presented so far, and having proceeded from an out-of-the-box GPT (see De Schryver 2025), to a customised GPT for lexicographic purposes, it is now necessary to take the next step for under-resourced languages, namely the fine-tuning of an LLM to (1) compile a dictionary, (2) act as a dictionary in its own right, or even (3) chat with the LLM as if it were a lexicographer. The differences between the three types are summarised in Table 1, adapted from De Schryver (2024b: 9).

Table 1: Comparing models for lexicography: Out-of-the-box GPT vs. customised GPT vs. fine-tuned GPT

Feature	Out-of-the-box GPT	Customised GPT	Fine-tuned GPT
Modifies core model	No	No	Yes
Training required	No	No, leverages pre-trained model	Yes, on a specific dataset
User input	Text prompt	Instructions and/or relevant documents	New data for the model to learn from
Complexity	Least complex, easy to use	Rather complex, potentially no-code/low-code	More complex, requires expertise
Cost	Least expensive	More expensive	A lot more expensive
Output control	Limited control over the generated text	More control over the direction of the output	High degree of control over the generated text
Suitability for 'exotic lexicography'	Limited, may not be familiar with 'exotic language'	Can be adapted to the 'exotic language' through prompts and instructions	More directly addresses the 'exotic language' through fine-tuning

The most important difference is listed in the first line of Table 1: one will need to modify the core model. In keeping with seeking a true challenge, as hinted at in the last line of Table 1: one could proceed with an exotic language about which the model has initially no clue. This is definitely a task for a future endeavour.⁹ But whatever the outcomes, they should be reported on, there and then.

PS.

Addendum C, available online (click [here](#)), gives an example of how the addition of the LLM Gemini to the search engine Google¹⁰ has turned that search engine into an intriguing GenAI chatbot. Search for 'broken uganda' today, and see how all data available on the Internet on this topic is first analysed and summarised, in real time, before listing the usual hyperlinks.

Endnotes

1. See <https://lmarena.ai/>.
2. See <https://pdev.org.uk/>.
3. See <https://www.collinsdictionary.com/dictionary/english/bank>, other verb uses.
4. See <https://www.collinsdictionary.com/dictionary/portuguese-english/banco>.
5. See <https://menhapublishers.com/dictionary/>.
6. Note that each token in an LLM like GPT-4o is either a whole word, part of a word, or punctuation, so a Lusoga corpus of 3.7 million word tokens has many more millions of LLM tokens.
7. Ironically, a massive effort to enlarge the Lusoga corpus was indeed undertaken starting at the end of June 2024, but after nearly three months of non-stop data recording and transcrip-

tion it was all in vain, as the material was stolen and the fieldworker himself suffered quite an impact in Uganda (De Schryver forthcoming-b).

8. More generally, and outside the field of lexicography, LLMs have (1) been trained to translate low-resource languages, such as for Inuktitut (40 thousand speakers; spoken in northern Canada) — Elsner and Needle (2023), or (2) been asked to learn new languages, such as for Zhuang (16 million speakers; spoken in southern China) — Zhang et al. (2024), or even (3) been taught languages with rare scripts, such as for Dzonghka (640,000 speakers; spoken in western Bhutan; Tibetan script), Santali (7.6 million speakers; spoken in India, Bangladesh and Nepal; Ol Chiki script), Nko (millions of speakers; spoken in West Africa; N'Ko script), Tamasheq (900,000 speakers; spoken in Mali and Burkina Faso; Tifinagh script), and Tigrinya (9.9 million speakers; spoken in Eritrea and Ethiopia; Ge'ez script) — Li et al. (2025).
9. That endeavour had actually been kickstarted, see Endnote 7. But the fieldworker was broken into, both literally (his hut was emptied) and sadly also literally (his skull and brain were broken up with machetes). After the fieldworker miraculously came fully back to life, he decided it was time to release his talk on 'Broken in Uganda', dating from August 2023.
10. See <https://www.google.com/>.
11. See <https://www.youtube.com/>.

References

- Barrett, G.** 2023. *Defin-o-Bots: Challenging A.I. to Create Usable Dictionary Content*. Paper presented at the 24th Biennial Conference of the Dictionary Society of North America, Boulder, CO, USA, 31 May–3 June 2023.
- De Schryver, G.-M.** 2023a. *Broken in Uganda: Uganda's National Curriculum Development Centre (NCDC)*. Paper presented at the Third Biannual Conference of the Language Association of Eastern Africa (LAEA), Makerere University, Kampala, Uganda, 15–16 August 2023, Kampala. <https://youtu.be/wYL6ZNIly5o>
- De Schryver, G.-M.** 2023b. *Contemporary Lexicography: A Case Study in AI [Keynote Lecture]*. Paper presented at the Inaugural Conference of the Association of Lexicography for the Americas: South, Central, Caribbean, and Mexico (Americalex-S), São Paulo, Brazil, 20–25 October 2023. <https://youtu.be/watch?v=yAxzTx7A2LU>
- De Schryver, G.-M.** 2024a. *Customising LLMs for Lexicography [Keynote Lecture]*. Paper presented at the 8th International Workshop on Cognitive Aspects of the Lexicon (CogALex-VIII), co-located with the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), Turin, Italy, 20 May 2024.
- De Schryver, G.-M.** 2024b. The Road towards Fine-tuned LLMs for Lexicography. Krek, S. (Ed.). 2024. *Book of Abstracts of the Workshop 'Large Language Models and Lexicography' @ EURALEX 2024, 8 October 2024, Cavtat, Croatia*: 6-11. Ljubljana: ELEXIS Association.
- De Schryver, G.-M.** 2025. Out-of-the-Box GPTs for Lexicography. *Lexikos* 35(2): 362-396.
- De Schryver, G.-M.** forthcoming-a. 100 Scholarly Echoes of Generative AI in Lexicography.
- De Schryver, G.-M.** forthcoming-b. Citing the Future: The Legacy of Patrick Hanks.
- De Schryver, G.-M. and M. Nabirye.** 2018. Corpus-driven Bantu Lexicography, Part 1: Organic Corpus Building for Lusoga. *Lexikos* 28: 32-78.

- De Schryver, G.-M. and M. Nabirye.** 2022. Towards a Monitor Corpus for a Bantu Language. A Case Study of Neology Detection in Lusoga. Klosa-Kückelhaus, A., S. Engelberg, C. Möhrs and P. Storjohann (Eds.). 2022. *Dictionaries and Society. Proceedings of the XX EURALEX International Congress, 12–16 July 2022, Mannheim, Germany*: 833-849. Mannheim: IDS-Verlag.
- Elsner, M. and J. Needle.** 2023. Translating a Low-resource Language Using GPT-3 and a Human-readable Dictionary. Nicolai, G., E. Chodroff, F. Mailhot and Ç. Çöltekin (Eds.). 2023. *Proceedings of the 20th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*: 1-13. Toronto: Association for Computational Linguistics.
- Grefenstette, G. and P. Hanks.** 2023. Competing Views of Word Meaning: Word Embeddings and Word Senses. *International Journal of Lexicography* 36(2): 211-219.
- Hanks, P.** 1979. Meaning and Grammar. Hanks, P. (Ed.), T.H. Long (Managing ed.), L. Urdang (Ed. director) and others (Eds.). 1979. *Collins Dictionary of the English Language*: xxxi-xxxv. London: William Collins Sons & Co.
- Hanks, P.** 1988. Typicality and Meaning Potentials. Snell-Hornby, M. (Ed.). 1988. *ZüriLEX '86 Proceedings: Papers read at the EURALEX International Conference, University of Zürich, 9–14 September 1986*: 37-47. Tübingen: A. Francke.
- Hanks, P.** 2000. Do Word Meanings Exist? *Computers and the Humanities* 34(1–2): 205-215.
- Hanks, P.** 2002. Mapping Meaning onto Use. Corréard, M.-H. (Ed.). 2002. *Lexicography and Natural Language Processing. A Festschrift in Honour of B.T.S. Atkins*: 156-198. s.l.: Euralex.
- Hanks, P.** 2004. Corpus Pattern Analysis. Williams, G. and S. Vessier (Eds.). 2004. *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004, Lorient, France, July 6–10, 2004*: 87-97. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud.
- Hanks, P.** 2008. Lexical Patterns: From Hornby to Hunston and Beyond. Bernal, E. and J. Decesaris (Eds.). 2008. *Proceedings of the XIII EURALEX International Congress (Barcelona, 15–19 July 2008)* (Sèrie Activitats 20): 89-129. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.
- Hanks, P.** 2012. How People Use Words to Make Meanings: Semantic Types Meet Valencies. Thomas, J. and A. Boulton (Eds.). 2012. *Input, Process and Product: Developments in Teaching and Language Corpora*: 54-69. Brno: Masaryk University Press.
- Hanks, P., I. El Maarouf and M. Oakes.** 2018. Flexibility of Multiword Expressions and Corpus Pattern Analysis. Sailer, M. and S. Markantonatou (Eds.). 2018. *Multiword Expressions: Insights from a Multilingual Perspective* (Phraseology and Multiword Expressions 1): 93-119. Berlin: Language Science Press.
- Hanks, P. and E. Franklin.** 2019. Do Online Resources Give Satisfactory Answers to Questions about Meaning and Phraseology? Corpas Pastor, G. and R. Mitkov (Eds.). 2019. *Computational and Corpus-based Phraseology. Third International Conference, EuroPhras 2019, Malaga, Spain, September 25–27, 2019, Proceedings* (Lecture Notes in Artificial Intelligence (LNAI), subseries of Lecture Notes in Computer Science (LNCS) 11755): 159-172. Berlin: Springer.
- Hanks, P. and W. Ma.** 2021. Meaning and Grammar in the Light of Corpus Pattern Analysis. *International Journal of Lexicography* 34(1): 135-149.
- Hanks, P. and S. Može.** 2019. The Way to Analyse 'way': A Case Study in Word-specific Local Grammar. *International Journal of Lexicography* 32(3): 247-269.
- Hanks, P. and J. Pustejovsky.** 2005. A Pattern Dictionary for Natural Language Processing. *Revue Française de linguistique appliquée* 10(2): 63-82.

- Li, Y., Z. Zhao and C. Scarton.** 2025. It's All about In-context Learning! Teaching Extremely Low-resource Languages to LLMs. Christodoulopoulos, C., T. Chakraborty, C. Rose and V. Peng (Eds.). 2025. *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*: 29532-29547. Suzhou: Association for Computational Linguistics.
- 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: Menha Publishers.
- Nabirye, M.** 2024. *Obuyumu mu Busoga [Entertainment in Busoga]* (History Series 3). Kampala: Menha Publishers.
- Nabirye, M. and G.-M. de Schryver.** 2013. Digitizing the Monolingual Lusoga Dictionary: Challenges and Prospects. *Lexikos* 23: 297-322.
- Nabirye, M., G.-M. de Schryver, D. Joffe and M.J. MacLeod.** 2012. *e-Eiwanika ly'Olusoga [Digital Lusoga Dictionary]* (Linguistics Series 3). Kampala / Cape Town: Menha Publishers / TshwaneDJe HLT.
- Zhang, C., X. Liu, J. Lin and Y. Feng.** 2024. Teaching Large Language Models an Unseen Language on the Fly. Ku, L.-W., A. Martins and V. Srikumar (Eds.). 2024. *Findings of the Association for Computational Linguistics: ACL 2024*: 8783-8800. Bangkok: Association for Computational Linguistics.

A Systematic Literature Review of Over Three Decades of Dictionary Use Research in *Lexikos*

Balázs Fajt, *Centre for Foreign Languages and Language Exams,
Budapest University of Economics and Business,
Budapest, Hungary*
(fajt.balazs@uni-bge.hu)
(<https://orcid.org/0000-0003-4983-3962>)

Mátyás Bánhegyi, *Centre for Foreign Languages and Language Exams,
Budapest University of Economics and Business, Budapest, Hungary*
(banhegyi.matyas@uni-bge.hu)
(<https://orcid.org/0000-0003-1011-0345>)

and

Katalin P. Márkus, *Department of English Linguistics, Károli Gáspár
University of the Reformed Church in Hungary, Budapest, Hungary*
(p.markus.kata@kre.hu)
(<https://orcid.org/0000-0001-6826-7631>)

Abstract: This study presents a systematic literature review of empirical research on dictionary use published in the journal *Lexikos* over its 35-volume history (1991–2025). The paper employs the PRISMA methodology. Based on 44 relevant studies identified, the review analyses trends in research focus, methodology, educational context, geographic distribution, and investigated languages. Findings reveal a growing interest in dictionary use, particularly since 2012, with a strong emphasis on English-language contexts and higher education. Methodologically, mixed-methods and cross-sectional designs dominate, while longitudinal research remains notably absent. Asia and Europe are the most represented regions, although studies span diverse global contexts. The review highlights underexplored areas such as primary education, non-English dictionary use, and longitudinal perspectives. The review also calls for future research that addresses these gaps, explores dictionary use at different language proficiency levels, and examines the pedagogical integration of digital dictionary tools. The findings provide valuable insights for researchers aiming to investigate dictionary use and dictionary-based learning.

Keywords: DICTIONARY USE, SYSTEMATIC REVIEW, USER BEHAVIOUR, PRISMA, METHODOLOGY, *LEXIKOS*, LANGUAGE EDUCATION, EMPIRICAL STUDIES, QUANTITATIVE RESEARCH, QUALITATIVE RESEARCH, MIXED METHODS RESEARCH

Zusammenfassung: Eine systematische Literaturübersicht über mehr als drei Jahrzehnte Wörterbuchgebrauchsforschung in *Lexikos*. In dieser Studie wird eine systematische Literaturübersicht über die empirische Forschung zum Wörterbuchgebrauch vorgestellt,

die in der Zeitschrift *Lexikos* im Laufe ihrer 35 Bände umfassenden Geschichte (1991–2025) veröffentlicht wurde. Die Studie basiert auf der PRISMA-Methode. Auf der Grundlage von 44 identifizierten relevanten Studien werden Trends in Bezug auf Forschungsschwerpunkte, Methodik, Bildungskontext, geografische Verteilung und untersuchte Sprachen analysiert. Die Ergebnisse zeigen ein wachsendes Interesse an der Wörterbuchnutzung, insbesondere seit 2012, mit einem starken Schwerpunkt auf englischsprachigen Kontexten und Hochschulbildung. Methodisch gesehen dominieren gemischte Methoden und Querschnittsstudien, während Längsschnittstudien nach wie vor nicht vorhanden sind. Asien und Europa sind die am stärksten vertretenen Regionen, obwohl die Studien verschiedene globale Kontexte umfassen. Der Bericht hebt bisher wenig erforschte Bereiche wie die Grundschulbildung, die Verwendung nicht-englischer Wörterbücher und die Längsschnittperspektive hervor. Die Übersicht fordert außerdem zukünftige Forschungsarbeiten, die diese Lücken schließen, den Wörterbuchgebrauch auf verschiedenen Sprachniveaus untersuchen und die pädagogische Integration von digitalen Wörterbuchwerkzeugen untersuchen. Die Ergebnisse liefern wertvolle Erkenntnisse für Forscher, die die Wörterbuchnutzung und das wörterbuchgestützte Lernen untersuchen wollen.

Stichwörter: WÖRTERBUCHBENUTZUNG, SYSTEMATISCHE ÜBERPRÜFUNG, BENUTZERVERHALTEN, PRISMA, METHODIK, *LEXIKOS*, SPRACHUNTERRICHT, EMPIRISCHE STUDIEN, QUANTITATIVE FORSCHUNG, QUALITATIVE FORSCHUNG, MIXED METHODS RESEARCH

1. Introduction

Dictionaries have long been regarded as indispensable tools in language learning and teaching that provide learners with access to a wealth of lexical, grammatical and cultural information (Karras 2019). Their importance spans various educational contexts from traditional classroom environments through digital learning to even self-directed learning scenarios (Verhun et al. 2024). Over the years, the study of dictionary use has evolved significantly, drawing the interest of researchers in fields such as second language acquisition (SLA), lexicography, and educational technology. Early studies focused primarily on the utilisation of print dictionaries, and investigated how users interacted with monolingual (Barnhart 1962, Quirk 1973), bilingual (Tomaszczyk 1979), and specialized dictionaries (Bergenholtz and Tarp 1995) to enhance their vocabulary (Bogaards 2010, Laufer and Hill 2000, Laufer 1993, Laufer and Hadar 1997), grammar (Bogaards and Van der Kloot 2001, Lew and Dziemianko 2006), reading comprehension (Ard 1982, Bensoussan et al. 1984, Knight 1994, Nesi and Meara 1991, Tono 2001), and writing skills (Bogaards 1996, Dziemianko 2010, Hartmann 2008, Laufer and Hadar 1997). These studies laid the groundwork for understanding the behaviours, strategies, and challenges learners face when using dictionaries as learning aids. With technological advancements, the focus of research has broadened to include electronic, digital and online dictionaries (Dziemianko 2010, Rundell 2015, Taylor and Chan 1994), which offer more dynamic and user-friendly interfaces compared to their print counterparts. These digital tools have introduced features such as audio pronunciations, visual aids, and context-sensitive

or fuzzy search options, thereby fundamentally changing the way users interact with lexical information. Consequently, research began to investigate the effectiveness of digital dictionaries, exploring their impact on language learning outcomes (Dziemianko 2012, Lew and Galas 2008), user preferences (Müller-Spitzer and Kopleinig 2014), and the cognitive processes involved (Miller 1986, Okuyama and Igarashi 2007) in dictionary consultation. The shift toward digital dictionaries has raised new questions about how learners access, interpret, and utilize the information provided (Bogaards 1998, Carstens 1995, Scholfield 1999), and this prompted ongoing investigations into the pedagogical implications of integrating dictionary use into language instruction (Chi 2003, Hartmann 1999, Lew and Galas 2008).

In this diverse research landscape, the study of dictionary use remains a dynamic field, reflecting the continuous evolution of both language learning practices and lexicographical innovations. As such, understanding how dictionaries are utilized in various learning contexts is relevant not only for language educators but also for lexicographers aiming to develop more effective reference tools (Atkins and Varantola 1997, Lew and De Schryver 2014, Miller 1986, Müller-Spitzer and Kopleinig 2014). The growing body of such research underscores the importance of conducting systematic literature reviews (e.g., Ge et al. 2024) and meta-analyses (e.g., Zhang et al. 2021) to map the current state of knowledge, identify trends, and highlight areas that require further investigation.

Within the field of linguistics, *Lexikos* has established itself over the past three decades as a leading forum for lexicographical scholarship. With 35 volumes published to date, the journal is set to commemorate its 35th anniversary in 2025 as a specialized and influential journal dedicated to the advancement of lexicographic research. It has consistently featured a diverse collection of studies focusing on dictionary use, including user preferences, lookup behaviours, and digital dictionary developments. Despite the existence of numerous reviews and meta-analyses on dictionary use, there is a noticeable gap in research that specifically examines the contributions of *Lexikos* to this body of knowledge. Therefore, the present study seeks to fill this gap by systematically reviewing the literature on dictionary use (and users) published in *Lexikos*. By focusing exclusively on this journal, this study aims to uncover trends, patterns, and areas that may warrant further exploration. Through utilizing the PRISMA methodology, this review synthesizes the findings from *Lexikos* in order to provide a detailed overview of lexicographical research conducted over the past three decades.

Methodology

The present study employs the Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) methodology (Page et al. 2021) to systematically identify, select, and synthesize the literature on dictionary use published in *Lexikos*. The methodology is comprised of several key steps, namely

defining inclusion and exclusion criteria, identifying sources, screening for eligibility, and analysing the data.

A comprehensive search was conducted in *Lexikos* on 15 April 2025: 35 volumes published from the journal's inception in 1991 to 2025 were scrutinised. Altogether there were 1028 electronic items (including articles, forewords, editorials, full volumes in a separate publication announcement, front matters and tables of contents), out of which 915 were articles. However, the search targeted (research) articles that investigated any aspect of dictionary use only, including but not limited to dictionary consultation behaviours, digital vs. print dictionary usage, the impact of dictionaries on language learning outcomes, user preferences, and the pedagogical implications of dictionary use. When searching for articles, we decided to manually examine each volume of *Lexikos* to ensure that no relevant studies were overlooked. This involved systematically going through the titles and abstracts of every article in each volume. By manually reviewing the content, we aimed to capture studies that could not be identified through keyword searches alone, thereby enhancing the thoroughness and accuracy of the study selection process.

Inclusion criteria were established to identify articles that specifically focused on dictionary use. The criteria were as follows:

1. published in *Lexikos*: only articles published in the journal *Lexikos* were included to ensure a focused analysis of research trends within the journal;
2. focus on dictionary use: studies must involve theoretical or empirical research related to any form of dictionary use, user behaviour, or dictionary-based language learning, including both qualitative and quantitative as well as mixed methods studies;
3. language: articles published in Afrikaans, Dutch, English, French or German (the language of contributions in *Lexikos*) were all included in the analysis.

Exclusion criteria were designed to eliminate studies that were not relevant to the review's objectives. Articles were excluded if their primary focus was not on dictionary use.

As explained earlier, the initial search identified all articles published in *Lexikos*. After that, relevant titles, abstracts and keywords of all the identified articles were screened for relevance based on the above inclusion and exclusion criteria. In cases where the title, abstract and keywords were insufficient to determine eligibility, the full text of the article was examined. The screening was conducted by all three authors to ensure consistency. Any disagreements regarding article eligibility were resolved through discussion until a consensus was reached. The results of the selection process were recorded in a PRISMA flow diagram (see Figure 1), detailing the number of articles identified, screened, assessed for eligibility, and included in the final review. This flow diagram provides transparency and replicability, as recommended by the PRISMA guidelines (Page et al. 2021).

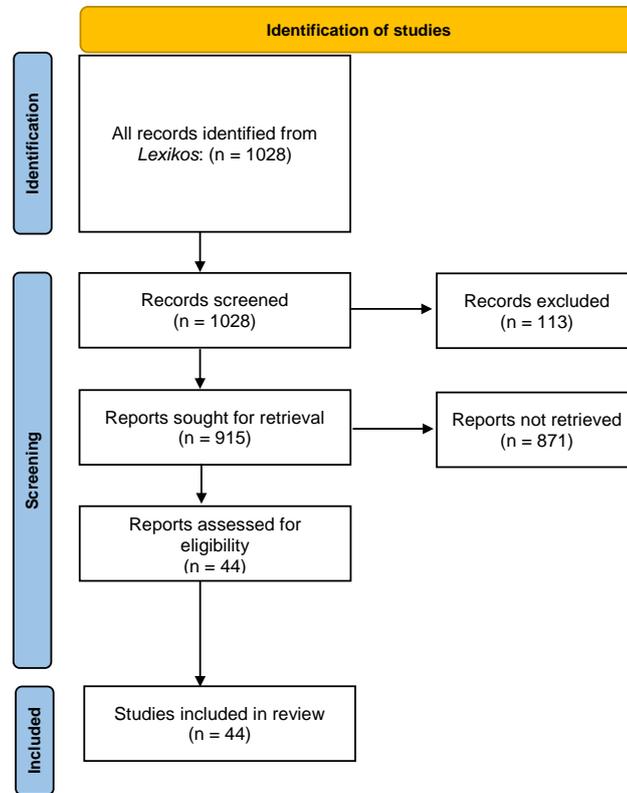


Figure 1

Data from the included studies were extracted systematically using the following coding sheet. Key information extracted from each study included:

- bibliographic details: author(s), year of publication, and volume number of *Lexikos*;
- research focus: the specific aspect(s) of dictionary use addressed in the study (e.g., user behaviour, etc.);
- study design and methodology: type of study (e.g., qualitative, quantitative, or mixed methods), number of participants, country from which the participants/respondents were drawn, languages investigated, sample size, and data collection methods.

The coding process was carried out by all authors to ensure accuracy and consistency. Any discrepancies in data extraction were discussed and resolved.

As a next step, a narrative synthesis approach was employed to analyse the findings of the studies included. Accordingly, the extracted data were organized thematically based on the research focus and methodology, which allowed for the identification of common trends, and evolving themes in dictionary use

over time. Descriptive statistics (e.g., n = number of items) were also used to summarize the frequency of different research topics, methodologies, and findings across the studies.

Finally, to ensure the robustness of the review, the quality of the included studies was assessed using a standardized checklist adapted to the context of dictionary use research. This checklist evaluated factors such as the clarity of research objectives, appropriateness of study design and methodology, transparency of data collection and analysis, and the validity of conclusions drawn. Studies were not excluded based on quality assessment.

Results

Regarding the number of published papers on dictionary use (n = 44, see Appendix), the analysis of publication trends — based on the corpus — clearly demonstrates evolving and growing scholarly interest in dictionary use studies over the examined time period (see Figure 2).

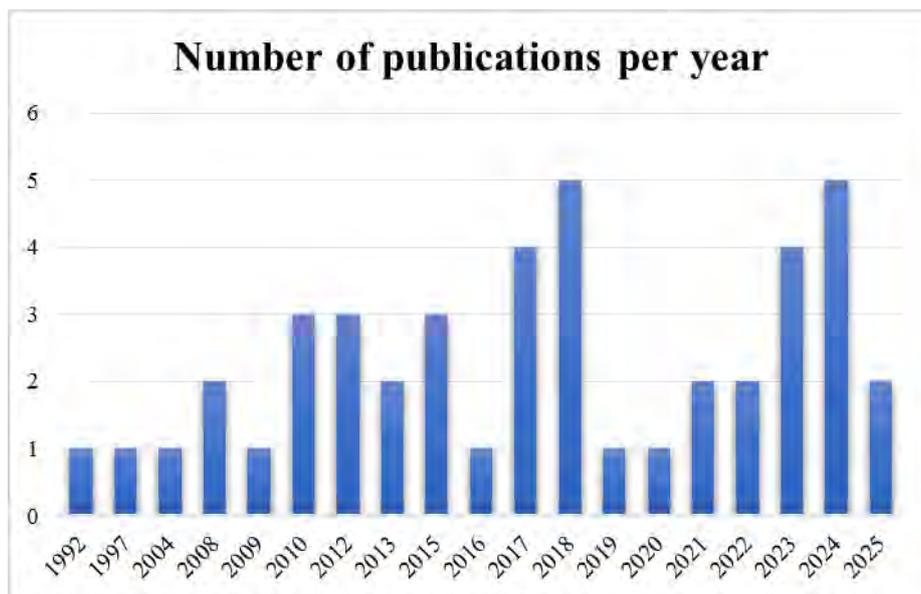


Figure 2

Initially, from 1992 to 2009, research papers in this field appeared sporadically and remained at a relatively low number typically ranging between one and two studies per year. A noticeable shift occurred in 2012, when the frequency of publications on dictionary use began to increase. This development coincided with the rapid spread of online dictionary use, supported by the growing accessi-

bility, speed, and reliability of internet connections. A particularly significant event was Macmillan's announcement in 2012 that, from 2013 onwards, it would cease publishing dictionaries in book form and instead would focus exclusively on digital resources (EURALEX 2012). This shift symbolized a broader trend: web-based dictionaries were becoming increasingly popular and dominant, opening up new research avenues, as the use of digital and online dictionaries involves fundamentally different practices compared to paper dictionaries.

The upward trajectory peaked in the years 2017 and 2018, which represented the highest number of scholarly contributions during this growth period. Following this peak, there was a temporary decrease in the rate of published works, probably due to the COVID-19 situation, which prevented the conducting of many studies building on formerly used research methodologies. However, the data reveal a recent surge in research activities, particularly in 2023 and 2024: this suggests renewed scholarly interest and potentially new directions or methodological advancements in the field.

The language distribution of the analysed studies further highlights a clear dominance of English: nearly all publications (43 out of 44) were written in English, and only one paper appeared in Afrikaans. This dominance underscores the status of English as the primary language of academic communication in the field of lexicographical and dictionary use research. Given that English is widely regarded as the global lingua franca as well as the language of science (O'Neil 2018), this overwhelming preference is hardly surprising. It suggests not only the international scope of the field but also the global scholarly community's sustained interest in issues related to dictionary usage. Moreover, the use of a global lingua franca, such as the English language, indubitably facilitates the wider dissemination and influence of these studies across diverse academic contexts.

From a research methodological point of view, the reviewed studies point to a distinct inclination towards certain research paradigms. However, before examining these in detail, it is important to acknowledge several previous works closely related to our topic. One notable systematic literature review is *Lexikos at Eighteen* by De Schryver (2009), which focused exclusively on articles published in *Lexikos*. More recently, Ge et al. (2024) conducted a scoping review of studies on dictionary use and language learning, offering a broader, systematic overview across various journals. In addition to these major reviews, two other studies have provided valuable insights into dictionary use in specific national contexts: Margalitadze and Meladze (2023) explored dictionary use in Georgia, while P. Márkus and Dringó-Horváth (2023) investigated the Hungarian context. Although these works are closely related to the topic of dictionary use, we did not include or categorize them within our dataset, as they are primarily descriptive in nature and/or synthesize findings from earlier studies, rather than present new empirical research.

The rest of papers may be classified into different categories. Mixed-methods designs are particularly prevalent, featuring in 12 studies, which signals a marked tendency among researchers to involve both qualitative and quantitative approaches for gaining a more holistic understanding of dictionary use (habits).

This preference is consistent with findings from broader review studies in the field. In a recent meta-review of mixed-methods research throughout the 2020s, Riazi and Farsani (2024) cite the work of Amini Farsani et al. (2021), who — drawing their conclusion on the basis of an analysis of 3,814 empirical articles published between 2009 and 2018 in 18 leading applied linguistics journals — found that, after quantitative studies (42.6%), mixed-methods research constituted the second most frequent methodological approach (25.9%). This growing reliance on mixed methods likely reflects the complex, context-dependent nature of dictionary use, which often requires both statistical generalisation and in-depth contextual insights.

Regarding the quantitative research paradigm, a total of 8 studies in our corpus were quantitative in nature. Notably, the dataset also includes four replication studies (Chen and Liu 2023, Ekwa Ebanéga and Moussavou 2008, Jewalani et al. 2019, Li et al. 2024). Although replication studies remain relatively rare in applied linguistics, their presence in our corpus reflects a growing emphasis on methodological rigour and the validation of previous findings. This trend aligns with the increasing recognition of the value of replication studies in the field. In fact, a recent special issue of *Studies in Second Language Acquisition* was dedicated entirely to the topic, highlighting the field's urgent need for more replication studies. As noted in the introduction to that issue (see McManus 2024), replication not only strengthens the empirical foundations of linguistic research but also contributes significantly to a deeper and more reliable understanding of key issues. On the other hand, 10 studies (Al-Ajmi 2008, Benzehra and McCreary 2010, Boonmoh 2012, Campoy-Cubillo 2021, Chaves-Yuste and Márquez-Cruz 2025, Gavriilidou et al. 2024, Kim 2017, Kim 2018, Milić et al. 2018, Pujol and Corrius 2013) explicitly utilized either a pre-test and post-test design or an experimental-group and control-group design. These designs allowed researchers to assess the immediate effects of dictionary use interventions on specific language learning outcomes, such as vocabulary acquisition, comprehension, or test performance.

Regarding purely qualitative research methods in the sample, 6 studies were identified. These studies employ various research methodologies with distinct features and purposes. Law (2024) uses qualitative content analysis to examine the effectiveness of secondary school EFL textbooks for dictionary use training. Furthermore, using the think-aloud protocol method, Vrbinc, Farina and Vrbinc (2022) compare dictionary look-up behaviours. In a descriptive study, Xue (2017) investigates how learner's dictionaries represent the cultural dimension of meaning for EFL learners. Through trends analysis De Schryver (2009) provides a comprehensive overview of *Lexikos'* contributions to key themes and developments in lexicography. Relying on log file analysis, De Schryver, Joffe, Joffe and Hillewaert (2006) study dictionary users' look-up behaviour. In addition, with the help of self-reflections and offering insights into dictionary use experience, Mdee (1997) compares university students' dictionary use and needs. The above studies present different approaches to lexicography research: qualitative content analysis and descriptive research are more systematic in nature, whereas think-

aloud protocols and self-reflections are more introspective. On the other hand, trends analysis and log file analysis are more focused on temporal changes and usage patterns, respectively.

An analysis of the applied research designs reveals clear methodological preferences and highlights notable research patterns. Cross-sectional studies seem dominant, appearing in 34 studies: these clearly indicate researchers' preference for gathering data at a single, specific point in time to gain immediate insights and to conduct comparative analyses of dictionary use. Interestingly, among the papers examined, no explicitly identified longitudinal studies were found. This gap supports Nesi's (2013) observation, who claims that most studies merely take a "snapshot" of dictionary user skills. This suggests that research which tracks participants and observes dictionary usage patterns over extended periods is notably absent from this scholarly domain. This enables the identification of future research directions.

Overall, the examined studies display considerable variability in participant numbers. The sample sizes range widely from small-scale studies involving as few as 10 or 13 participants (generally qualitative studies) to large-scale research with over 300 participants. More specifically, small sample sizes (fewer than 50 participants) are relatively common, with several studies ($n=11$) reporting groups of 10–45 participants, and a moderate number of studies ($n=15$) involved between 50 and 150 participants. Finally, large-scale studies involving more than 150 participants are less frequent but notable ($n=8$), with reported sample sizes of between 168 and 509 participants. This distribution highlights two trends: while small to moderately sized context-specific studies dominate, there is a noteworthy presence of large-sample research, which may allow for more robust statistical analyses and greater generalizability. Nevertheless, the overall variability in sample sizes suggests a lack of standardization in research designs within the field.

The analysis of educational levels targeted by the reviewed studies reveals distinct trends and research preferences. Studies clearly prioritise higher education contexts, with 31 studies explicitly focusing on university-level learners. This suggests that dictionary research predominantly explores issues pertinent to adult learners. This might be explained by easier access to participants or the importance of dictionary use at advanced proficiency levels (cf. Béjoint 1981, Battenburg 1989, Kosem 2008, Nesi 2000). On the other hand, research explicitly focusing on younger learners is noticeably scarce. Only two studies examined dictionary use at the primary education level: one involved mainstream and intercultural primary school pupils aged 10–12 (grades 5–6) (Gavriilidou et al. 2024), while the other study (Van der Merwe 2012) concentrated on primary learners in Grade 4. Additionally, one study extended the scope of research across educational levels: a study by Margalitadze and Meladze (2023) examined dictionary use across primary, secondary and higher education levels, and another focused concurrently on secondary and higher education (Tono 1992).

An analysis of the countries from which the participants/respondents of the analysed studies were drawn highlights distinct geographic areas of research.

China stands out as the most frequently examined context, with seven studies. Following China, Poland appears prominent with four studies, which indicates focused and consistent attention on dictionary use in Polish contexts. Thailand and South Africa each feature in three studies, and a moderate level of interest is apparent in countries like Spain ($n = 3$), Indonesia ($n = 2$), and Korea ($n = 2$), each represented by two studies. At a continental comparison, Asia emerges as the most represented region, with countries such as China, Thailand, Indonesia, Korea, Japan, Taiwan, Jordan, and Kuwait (with the latter two countries classified under 'Asia' in the current study) contributing to the body of research. **Europe** also shows substantial representation, led by Poland and Spain, and supplemented by individual studies from Germany, Greece, Hungary, Serbia, and Slovenia. This points to a well-established tradition of lexicographic inquiry within the European academic landscape. Furthermore, Africa, while less represented overall, still shows a notable presence with three studies from Republic of South Africa and one each from Zimbabwe and Namibia. This suggests growing scholarly engagement with dictionary use on the African continent, particularly in multilingual or second-language learning contexts. It is also important to note that *Lexikos* — the journal analysed in this study — is the only journal in Africa exclusively devoted to lexicography, which likely contributes to the relatively higher visibility of African research in this field. North America and Oceania are less prominently featured, with the United States being the only representative from North America, and no studies originating from Oceania in the reviewed sample. South America is absent: this points to potential gaps in regional representation and opportunities for future research.

The analysis of languages investigated in the reviewed studies underscores the clear dominance of English, reflecting its global status and importance in educational and lexicographical research. English was the primary focus in 26 studies, which indicates widespread scholarly attention towards understanding dictionary use in English-language contexts (cf. Nied Curcio 2022). This focus on English is likely to be due to the international relevance of English and its role as a lingua franca and a language and medium of instruction. Beyond the English language, several studies explored bilingual or multilingual dictionary use. These studies investigated dictionary interactions between English and other languages including e.g. Serbian, Chinese, Arabic, Japanese, German, French, and Swahili. These studies reflect an interest in examining dictionary use in linguistically diverse settings and highlight cross-linguistic issues, challenges and opportunities. In addition, studies examined dictionary use related solely to specific languages like Greek, Georgian, Bahasa Indonesian, Afrikaans, Ndebele, and Gabonese. Though individually less represented, these studies illustrate curiosity about dictionary use concerning languages beyond the dominant global language of English.

Conclusion

This systematic literature review set out to examine dictionary use research published in *Lexikos*, a specialized journal of lexicography that has played a

pivotal role in shaping the discourse around dictionary-related studies over the past three decades. By analysing 44 empirical papers from 1991 to 2025 using the PRISMA framework, the review has revealed clear trends, methodological patterns, and emerging gaps in the field. The findings demonstrate that scholarly interest in dictionary use has grown significantly over time, with a notable increase in publications from 2012 onwards, coinciding with the rise of digital and online dictionaries. This shift has brought about methodological diversification, particularly in the increased use of mixed-methods designs. Despite this methodological expansion, cross-sectional research seems to dominate the field and there is scarcity of longitudinal research on dictionary use. In terms of research contexts, the field shows a focus on higher education, with minimal attention given to primary and secondary school learners. Geographically, the reviewed studies predominantly originate from Asia and Europe, with a strong emphasis on English-language contexts, both in terms of the language of publication and the language of investigation.

Importantly, the review identifies a noticeable lack of replication studies and large-scale investigations, as well as insufficient reporting of participant numbers in some cases, which are factors that can limit the generalizability and transparency of findings. Additionally, while the reviewed studies cover a wide range of topics such as user behaviour, digital innovations, and pedagogical applications, there remains untapped potential for deeper exploration into cognitive processes, long-term learning effects, and classroom-based interventions. As a final note, we encourage future authors in this field to place more emphasis on the consistent use of research methodology and terminology, as clearer conceptual frameworks and methodological transparency will strengthen the validity and impact of dictionary use research.

References

- Al-Ajmi, H.** 2008. The Effectiveness of Dictionary Examples in Decoding: The Case of Kuwaiti Learners of English. *Lexikos* 18: 15-26.
<https://doi.org/10.5788/18-0-474>
- Alzi'abi, S.E.** 2017. Guessing Verb-Adverb Collocations: Arab EFL Learners' Use of Electronic Dictionaries. *Lexikos* 27: 50-77.
<https://doi.org/10.5788/27-1-1394>
- Amini Farsani, M., H.R. Jamali, M. Beikmohammadi, B. Daneshvar Ghorbani and L. Soleimani.** 2021. Methodological Orientations, Academic Citations, and Scientific Collaboration in Applied Linguistics: What Do Research Synthesis and Bibliometrics Indicate? *System* 100(3): 102547.
<https://doi.org/10.1016/j.system.2021.102547>
- Ard, J.** 1982. The Use of Bilingual Dictionaries by ESL Students While Writing. *ITL Review of Applied Linguistics* 58: 1-27.
- Atkins, B.T.S. and K. Varantola.** 1997. Monitoring Dictionary Use. *International Journal of Lexicography* 10(1): 1-45.
- Barnhart, C.** 1962. Problems in Editing Commercial Monolingual Dictionaries. Householder, F.W. and S. Saporta (Eds.). 1962. *Problems in Lexicography*: 161-181. Bloomington: Indiana University Press.

- Battenburg, J.D.** 1989. *A Study of English Monolingual Learners' Dictionaries and their Users*. West Lafayette, Indiana, USA: Purdue University.
- Béjoint, H.** 1981. The Foreign Student's Use of Monolingual English Dictionaries: A Study of Language Needs and Reference Skills. *Applied Linguistics* II(3): 207-222.
- Bensoussan, M., D. Sim and R. Weiss.** 1984. The Effect of Dictionary Usage on EFL Test Performance Compared with Student and Teacher Attitudes and Expectations. *Reading in a Foreign Language* 2(2): 262-276.
- Benzehra, R. and D.R. McCreary.** 2010. The OEAD: New Perspectives on English–Arabic Dictionaries for English-Speaking Users and Users' Surveys. *Lexikos* 20: 1-25.
<https://doi.org/10.5788/20-0-130>
- Bergenholtz, H. and S. Tarp (Eds.).** 1995. *Manual of Specialised Lexicography. The Preparation of Specialised Dictionaries*. Amsterdam/Philadelphia: John Benjamins.
- Beyer, H.L. and A.K. Faul.** 2010. Namibian University Entrants' Concepts of 'a Dictionary'. *Lexikos* 20: 644-679.
<https://doi.org/10.5788/20-0-160>
- Bogaards, P.** 1996. Dictionaries for Learners of English. *International Journal of Lexicography* 9(4): 277-320.
- Bogaards, P.** 1998. Using Dictionaries: Which Words Are Looked Up by Foreign Language Learners? Atkins, B.T.S. and K. Varantola (Eds.). 1998. *Studies of Dictionary Use by Language Learners and Translators*: 151-157. Tübingen: Niemeyer.
- Bogaards, P.** 2010. Dictionaries and Second Language Acquisition. Dykstra, A. and T. Schoonheim (Eds.). 2010. *Proceedings of the XIV EURALEX International Congress, 6–10 July 2010, Leeuwarden*: 99-123. Ljouwert: Fryske Akademy.
- Bogaards, P. and W.A. van der Kloot.** 2001. The Use of Grammatical Information in Learners' Dictionaries. *International Journal of Lexicography* 14(2): 97-121.
<https://doi.org/10.1093/ijl/14.2.97>
- Boonmoh, A.** 2010. The Use of Pocket Electronic Dictionaries by Thai University Students. *Lexikos* 20: 680-690.
<https://doi.org/10.5788/20-0-161>
- Boonmoh, A.** 2012. E-dictionary Use under the Spotlight: Students' Use of Pocket Electronic Dictionaries for Writing. *Lexikos* 22: 43-68.
<https://doi.org/10.5788/22-1-997>
- Boonmoh, A.** 2021. Use of Dictionaries and Online Tools for Reading by Thai EFL Learners in a Naturalistic Setting. *Lexikos* 31(1): 239-258.
<https://doi.org/10.5788/31-1-1645>
- Çağlayan Özkan, A.** 2025. The Effectiveness of the "Real-Life-Based School Dictionary" as an Electronic Dictionary. *Lexikos* 35(1): 143-156.
<https://doi.org/10.5788/35-1-2006>
- Campoy-Cubillo, M.C.** 2021. Fostering Learners' Online Dictionary Skills through Active Dictionary Rubrics. *Lexikos* 31(1): 487-510.
<https://doi.org/10.5788/31-1-1663>
- Carstens, A.** 1995. Language Teaching and Dictionary Use: An Overview. *Lexikos* 5: 105-116.
- Chan, A.Y.W.** 2012. Advanced Cantonese ESL Learners' Use of a Monolingual Dictionary for Language Production. *Lexikos* 22: 109-138.
<https://doi.org/10.5788/22-1-1000>

- Chan, A.Y.W.** 2017. The Effectiveness of Using a Bilingualized Dictionary for Determining Noun Countability and Article Selection. *Lexikos* 27: 183-213.
<https://doi.org/10.5788/27-1-1399>
- Chaves-Yuste, B. and M. Márquez-Cruz.** 2025. Digital Lexicography in ESP: Enhancing Vocabulary Acquisition and Reading Comprehension in Students of Commerce Degree. *Lexikos* 35(1): 109-128.
<https://doi.org/10.5788/35-1-2004>
- Chen, Y. and S. Liu.** 2023. A Further Look into the Use of a Dictionary APP in EFL Writing: A Replication Study. *Lexikos* 33(1): 324-349.
<https://doi.org/10.5788/33-1-1820>
- Chi, M.L.A.** 2003. *An Empirical Study of the Efficacy of Integrating the Teaching of Dictionary Use into a Tertiary English Curriculum in Hong Kong. Vol. IV. Research Reports* (Ed. G. James). Hong Kong: Language Centre, Hong Kong University of Science and Technology, Language Center.
- Claassen, D.** 2015. 'n Evaluasie van *Pharos Aanleerderwoordeboek vir skole/Learner's Dictionary for Schools* en *Longman-HAT Afrikaans Dictionary and Grammar for English Speakers* in 'n Afrikaanse taalverwerwingskursus op universiteitsvlak. *Lexikos* 25: 31-55.
<https://doi.org/10.5788/25-1-1290>
- De Schryver, G.-M.** 2009. *Lexikos* at Eighteen: An Analysis. *Lexikos* 19: 372-403.
<https://doi.org/10.5788/19-0-445>
- De Schryver, G.-M., D. Joffe, P. Joffe and S. Hillewaert.** 2006. Do Dictionary Users Really Look Up Frequent Words? — On the Overestimation of the Value of Corpus-based Lexicography. *Lexikos* 16: 67-83.
<https://doi.org/10.5788/16-0-648>
- Dziemianko, A.** 2010. Paper or Electronic? The Role of Dictionary Form in Language Reception, Production and the Retention of Meaning and Collocations. *International Journal of Lexicography* 23(3): 257-273.
- Dziemianko, A.** 2012. On the Use(fulness) of Paper and Electronic Dictionaries. Granger, S. and M. Paquot (Eds.). 2012. *Electronic Lexicography*: 319-341. Oxford: Oxford University Press.
- Ekwa Ebanéga, G.-M. and F.T. Moussavou.** 2008. A Survey of the Dictionary Use of Gabonese Students at Two South African Universities. *Lexikos* 18: 349-365.
<https://doi.org/10.5788/18-0-491>
- EURALEX.** 2012. *End of Print Dictionaries at Macmillan*.
<https://euralex.org/2012/11/05/end-of-print-dictionaries-at-macmillan/> [10 April 2025].
- Fajt, B.** 2024. Exploring Dictionary Preferences: A Comparative Study of EFL and GFL Learners in Hungarian Higher Education. *Lexikos* 34(1): 419-436.
<https://doi.org/10.5788/34-1-1945>
- Fallianda, F.** 2020. A Survey of Indonesian Students' Use of Dictionaries. *Lexikos* 30(1): 609-628.
<https://doi.org/10.5788/30-1-1607>
- Gavriilidou, Z., A. Markos and E. Konstantinidou.** 2024. How Can We Raise Strategic Dictionary Use in the Classroom: The Effect of a Dictionary Awareness Program on Dictionary Use Strategies. *Lexikos* 34(1): 99-122.
<https://doi.org/10.5788/34-1-1882>
- Ge, X., S. Zhang, H. Xu and X. Zhang.** 2024. A Scoping Review of Studies into Dictionary Use and Language Learning. *Lexikos* 34(1): 331-354.
<https://doi.org/10.5788/34-1-1940>

- Hadebe, S.** 2004. Improving Dictionary Skills in Ndebele. *Lexikos* 14: 89-104.
<https://doi.org/10.5788/14-0-684>
- Hartmann, R.R.K. (Ed.)**. 1999. *Dictionaries in Language Learning. Recommendations, National Reports and Thematic Reports from the TNP Sub-project 9: Dictionaries*. Thematic Network Project in the Area of Languages, Berlin Free University.
- Hartmann, R.R.K.** 2008. Twenty-five Years of Dictionary Research: Taking Stock of Conferences and Other Lexicographic Events Since LEXeter '83. Bernal, E. and J. DeCesaris (Eds.). 2008. *Proceedings of the XIII EURALEX International Congress, 15-19 July 2008, Barcelona, Spain*: 131-148. Barcelona: Universitat Pompeu Fabra.
- Jewalani, A.P., M. Laksman-Huntley and H.Y. Anjarningsih.** 2019. Lexical Understanding of Native Bahasa Indonesia Speakers through Word Association to Improve Dictionary Definitions. *Lexikos* 29: 36-54.
<https://doi.org/10.5788/29-1-1511>
- Karras, I.** 2019. The Use of Dictionaries in English as Foreign Language Teaching and Learning. Osuchowska, D. and L. Harmon (Eds.). 2019. *Contents, Use, Usability: Dictionaries from the Perspective of a Translator and a Language Teacher*: 49-62. Bern: Peter Lang International Academic Publishing.
- Kim, S.** 2017. "Minimum Input, Maximum Output, Indeed!" Teaching Collocations Through Collocation Dictionary Skills Development. *Lexikos* 27: 265-286.
<https://doi.org/10.5788/27-1-1403>
- Kim, S.** 2018. A Lexicographic Approach to Teaching the English Article System: Help or Hindrance? *Lexikos* 28: 196-220.
<https://doi.org/10.5788/28-1-1462>
- Knight, S.** 1994. Dictionary Use While Reading: The Effects on Comprehension and Vocabulary Acquisition for Students of Different Verbal Abilities. *The Modern Language Journal* 78(3): 285-299.
<https://doi.org/10.2307/330108>
- Kosem, I.** 2008. Dictionaries for University Students: A Real Deal or Merely a Marketing Ploy? Bernal, E. and J. DeCesaris (Eds.). 2008. *Proceedings of the XIII EURALEX International Congress, 15-19 July 2008, Barcelona, Spain*: 1575-1584. Barcelona: Universitat Pompeu Fabra.
- Laufer, B.** 1993. The Effect of Dictionary Definitions and Examples on the Use and Comprehension of New L2 Words. *Cahiers de Lexicologie* 63(2): 131-142.
- Laufer, B. and L. Hadar.** 1997. Assessing the Effectiveness of Monolingual, Bilingual, and "Bilingualized" Dictionaries in the Comprehension and Production of New Words. *The Modern Language Journal* 81(2): 189-196.
- Laufer, B. and M. Hill.** 2000. What Lexical Information Do L2 Learners Select in a CALL Dictionary and How Does It Affect Word Retention? *Language Learning & Technology* 3(2): 58-76.
- Law, Wai-on.** 2024. Dictionary Use Training in Secondary School EFL Textbooks in Taiwan. *Lexikos* 34(1): 77-98.
<https://doi.org/10.5788/34-1-1880>
- Lew, R. and G.-M. de Schryver.** 2014. Dictionary Users in the Digital Revolution. *International Journal of Lexicography* 27(4): 341-359.
- Lew, R. and A. Dziemianko.** 2006. A New Type of Folk-inspired Definition in English Monolingual Learners' Dictionaries and its Usefulness for Conveying Syntactic Information. *International Journal of Lexicography* 19(3): 225-242.

- Lew, R. and K. Galas.** 2008. Can Dictionary Skills Be Taught? The Effectiveness of Lexicographic Training for Primary School Level Polish Learners of English. Bernal, E. and J. DeCesaris (Eds.). 2008. *Proceedings of the XIII EURALEX International Congress, 15-19 July 2008, Barcelona, Spain*: 1273-1285. Barcelona: Universitat Pompeu Fabra.
- Lew, R., M. Grzelak and M. Leszkowicz.** 2013. How Dictionary Users Choose Senses in Bilingual Dictionary Entries: An Eye-Tracking Study. *Lexikos* 23: 228-254.
<https://doi.org/10.5788/23-1-1213>
- Li, L. and H. Xu.** 2015. Using an Online Dictionary for Identifying the Meanings of Verb Phrases by Chinese EFL Learners. *Lexikos* 25: 191-209.
<https://doi.org/10.5788/25-1-1295>
- Li, L., H. Wang and H. Xu.** 2024. Adaptation and Validation of the Strategy Inventory for Electronic Dictionary Use (S.I.E.D.U.) for Chinese Learners. *Lexikos* 34(1): 245-268.
<https://doi.org/10.5788/34-1-1914>
- Liang, P. and D. Xu.** 2018. An Empirical Study of EFL Learners' Dictionary Use in Chinese-English Translation. *Lexikos* 28: 221-244.
<https://doi.org/10.5788/28-1-1463>
- Margalitadze, T. and G. Meladze.** 2023. How to Solve Problems in Dictionary Use: The Georgian Case. *Lexikos* 33(2): 63-78.
<https://doi.org/10.5788/33-2-1840>
- McManus, K.** 2024. Replication Studies in Second Language Acquisition Research: *Definitions, Issues, Resources, and Future Directions: Introduction to the Special Issue. Studies in Second Language Acquisition* 46(5): 1299-1319.
<https://doi.org/10.1017/S0272263124000652>
- Mdee, J.S.** 1997. Language Learner's Use of a Bilingual Dictionary: A Comparative Study of Dictionary Use and Needs. *Lexikos* 7: 94-106.
<https://doi.org/10.5788/7-1-974>
- Milić, M., T. Glušač and A. Kardoš.** 2018. The Effectiveness of Using Dictionaries as an Aid for Teaching Standardization of English-based Sports Terms in Serbian. *Lexikos* 28: 262-286.
<https://doi.org/10.5788/28-1-1465>
- Miller, G.A.** 1986. Dictionaries in the Mind. *Language and Cognitive Processes* 1(3): 171-185.
<https://doi.org/10.1080/01690968608407059>
- Müller-Spitzer, C. and A. Kopleinig.** 2014. Online Dictionaries: Expectations and Demands. Müller-Spitzer, C. (Ed.). 2014. *Using Online Dictionaries*. Lexicographica Series Maior 145: 143-188. Berlin: Walter de Gruyter.
- Müller-Spitzer, C., M.J. Domínguez Vázquez, M. Nied Curcio, I.M. Silva Dias and S. Wolfer.** 2018. Correct Hypotheses and Careful Reading Are Essential: Results of an Observational Study on Learners Using Online Language Resources. *Lexikos* 28: 287-315.
<https://doi.org/10.5788/28-1-1466>
- Nesi, H.** 2000. *The Use and Abuse of EFL Dictionaries: How Learners of English as a Foreign Language Read and Interpret Dictionary Entries*. Tübingen: Max Niemeyer.
- Nesi, H.** 2013. Researching Users and Uses of Dictionaries. Jackson, H. (Ed.). 2013. *The Bloomsbury Companion to Lexicography*: 62-74. London: Bloomsbury Academic.
- Nesi, H. and P. Meara.** 1991. How Using Dictionaries Affects Performance in Multiple-choice EFL Tests. *Reading in a Foreign Language* 8(1): 631-643.

- Nied Curcio, M.** 2022. Dictionaries, Foreign Language Learners and Teachers. New Challenges in the Digital Era. Klosa-Kückelhaus, A., S. Engelberg, Ch. Möhrs and P. Storzjohann (Eds.). 2022. *Dictionaries and Society: Proceedings of the XX EURALEX International Congress, 12–16 July 2022, Mannheim, Germany*: 71-84. Mannheim: IDS-Verlag.
- Okuyama, Y. and H. Igarashi.** 2007. Think-aloud Protocol on Dictionary Use by Advanced Learners of Japanese. *The JALT CALL Journal* 3(1–2): 45-58.
- O'Neil, D.** 2018. English as the Lingua Franca of International Publishing. *World Englishes* 37(2): 146-165.
- Page, M.J., J.E. McKenzie, P.M. Bossuyt, I. Boutron, T. Hoffmann, C.D. Mulrow et al.** 2021. The PRISMA 2020 Statement: An Updated Guideline for Reporting Systematic Reviews. *BMJ* 372(71): 1-9.
<https://doi.org/10.1136/bmj.n71>
- P. Márkus, K. and I. Dringó-Horváth.** 2023. Developing Dictionary Skills through Monolingual and Bilingual English Dictionaries at Tertiary-level Education in Hungary. *Lexikos* 33(1): 350-381.
<https://doi.org/10.5788/33-1-1821>
- Ptasznik, B.** 2022. After the Digital Revolution: Dictionary Preferences of English Majors at a European University. *Lexikos* 32(1): 220-249.
<https://doi.org/10.5788/32-1-1721>
- Ptasznik, B.** 2023. Evaluating the Usefulness of the Learning Tools in Monolingual Online Dictionaries for Learners of English: Gauging the Preferences of Polish Students of English. *Lexikos* 33(1): 109-124.
<https://doi.org/10.5788/33-1-1799>
- Ptasznik, B. and R. Lew.** 2018. Which Learning Tools Accompanying the Paid Online Version of LDOCE Do Advanced Learners of English Find Useful? *Lexikos* 28: 448-460.
<https://doi.org/10.5788/28-1-1474>
- Pujol, D. and M. Corrius.** 2013. The Role of the Learner's Native Culture in EFL Dictionaries: An Experimental Study. *Lexikos* 23: 565-584.
<https://doi.org/10.5788/23-1-1230>
- Quirk, R.** 1973. The Social Impact of Dictionaries in the UK. McDavid, R.I. and A.R. Duckert (Eds.). 1973. *Lexicography in English*: 76-88. New York: New York: Academy of Sciences.
- Riazi, A.M. and M.A. Farsani.** 2024. Mixed-methods Research in Applied Linguistics: Charting the Progress through the Second Decade of the Twenty-first Century. *Language Teaching* 57(2): 143-182.
<https://doi.org/10.1017/S0261444823000332>
- Rundell, M.** 2015. From Print to Digital: Implications for Dictionary Policy and Lexicographic Conventions. *Lexikos* 25: 301-322.
<https://doi.org/10.5788/25-1-1301>
- Schierholz, S.J.** 2015. Methods in Lexicography and Dictionary Research. *Lexikos* 25: 323-352.
<https://doi.org/10.5788/25-1-1302>
- Scholfield, P.** 1999. Dictionary Use in Reception. *International Journal of Lexicography* 12(1): 13-34.
- Taylor, A. and A. Chan.** 1994. Pocket Electronic Dictionaries and Their Use. Martin, W., W. Meijs, M. Moerland, E. ten Pas, P. van Sterkenburg and P. Vossen (Eds.). 1994. *EURALEX '94 Proceedings. Papers Submitted to the 6th EURALEX International Congress on Lexicography in Amsterdam, The Netherlands*: 598-605. Amsterdam: EURALEX.

- Tomaszczyk, J.** 1979. Dictionaries: Users and Uses. *Glottodidactica* 12: 103-119.
- Tono, Y.** 1992. The Effect of Menus on EFL Learners' Look-up Processes. *Lexikos* 2: 230-253.
<https://doi.org/10.5788/2-1-1136>
- Tono, Y.** 2001. *Research on Dictionary Use in the Context of Foreign Language Learning: Focus on Reading Comprehension*. Lexicographica Series Maior 106. Tübingen: Max Niemeyer.
- Van der Merwe, M.F.** 2012. A Study of the Use of the *HAT Afrikaanse Skoolwoordeboek* by Primary School Children. *Lexikos* 22: 352-366.
<https://doi.org/10.5788/22-1-1012>
- Verhun, T., M. Davydovych, K. Parnus, M. Kosheleva, I. Biliak and O. Khorosh.** 2024. Algorithms for Optimizing Vocabulary Acquisition in Language Learning. *Proceedings of the International Conference on Applied Innovation in IT* 12(1): 105-114.
- Vrbinc, M., D.M.T.Cr. Farina and A. Vrbinc.** 2022. Who is Sharpest at Looking Up *sharp*? Comparing Two Parallel Groups of Dictionary Users. *Lexikos* 32(1): 368-391.
<https://doi.org/10.5788/32-1-1750>
- Xue, M.** 2017. Representing the Cultural Dimension of Meaning in Learner's Dictionaries — From the Perspective of Chinese EFL Learners in L2 Reception. *Lexikos* 27: 578-596.
<https://doi.org/10.5788/27-1-1416>
- Zhang, S., H. Xu and X. Zhang.** 2021. The Effects of Dictionary Use on Second Language Vocabulary Acquisition: A Meta-analysis. *International Journal of Lexicography* 34(1): 1-38.
<https://doi.org/10.1093/ijl/ecaa010>
- Zou, D.** 2016. Comparing Dictionary-induced Vocabulary Learning and Inferencing in the Context of Reading. *Lexikos* 26: 372-390.
<https://doi.org/10.5788/26-1-1345>

APPENDIX

Paper	Research design and number of participants (n)	Target group	Country	Language investigated
Chaves-Yuste, B. and M. Márquez-Cruz (2025)	experiment (n=90)	first year university students	Spain	English
Çağlayan Özkan, A. (2025)	quantitative (questionnaire + achievement test) (n=83)	secondary school students (5th and 7th graders)	Turkey	Turkish
Ge, X., S. Zhang, H. Xu and X. Zhang (2024)	scoping review	—	—	—
Fajt, B. (2024)	quantitative (n=371)	English and German majors at university	Hungary	English, German
Li, L., H. Wang and H. Xu (2024)	quantitative (n=494)	English majors in their third year (probably around C1)	China	English
Gavriilidou, Z., A. Markos and E. Konstantinidou (2024)	quantitative (n=150)	primary school pupils (aged 10–12) (native)	Greece	Greek
Law, Wai-on (2024)	qualitative content analysis	—	Taiwan	English
Margalitadze, T. and G. Meladze (2023)	analyses the situation of dictionary use in Georgia	Primary, secondary and higher education	Georgia	—
P. Márkus, K. and I. Dringó-Horváth (2023)	analyses the situation of dictionary use in Hungary	—	Hungary	—
Chen, Y. and S. Liu (2023)	mixed methods (n=62)	English majors (C1)	China	English
Ptasznik, B. (2023)	mixed methods (n=318)	English majors (C1–C2)	Poland	English
Vrbinc, M., D.M.T.Cr. Farina and A. Vrbinc (2022)	qualitative (n=17) (think-aloud protocol)	English majors (C1–C2)	Slovenia	English
Ptasznik, B. (2022)	mixed methods (n=168)	English majors (B2–C1)	Poland	English
Campoy-Cubillo, M.C. (2021)	mixed methods (n=75)	English majors (B1)	Spain	English
Boonmoh, A. (2021)	experiment (?) in a case study (n=14)	university students (A2–B1)	Thailand	English
Fallianda, F. (2020)	quantitative (n=88)	university students (A2–B1)	Indonesia	English

Jewelani, A.P., M. Laksman-Huntley, and H.Y. Anjarningsih (2019)	observational study, experiment (n=45)	native speakers	Indonesia	Bahasa Indonesia
Ptasznik, B. and R. Lew (2018)	experiment (n=114)	university students (B2–C1)	Poland	English
Müller-Spitzer, C., M.J. Domínguez Vázquez, M. Nied Curcio, I.M. Silva Dias and S. Wolfer (2018)	observational study (mixed methods: different quantitative and qualitative methods via the examination) (n=42)	university students (A2–B1)	Germany, but the students were from Portugal, Italy and Spain	English
Milić, M., T. Glušac and A. Kardoš (2018)	quantitative (n=255)	university students (B1)	Serbia	Serbian, English
Liang, P. and D. Xu (2018)	mixed methods (n=50) (observational and qualitative)	university students (B1–B2)	China	Chinese, English
Kim, S. (2018)	quasi experiment (n=43)	university students (B1–B2)	Korea	English
Xue, M. (2017)	descriptive (qualitative)	—	China	—
Kim, S. (2017)	experiment (n=59)	university students (A2–C1)	Korea	English
Chan, A.Y.W. (2017)	mixed methods (n=30) (quantitative + test)	university students (B2–C1)	China	English
Alzi'abi, S.E. (2017)	mixed methods (n=82)	university students (B2–C1)	Jordan	English
Zou, D. (2016)	mixed methods (n=104) (experiment + qualitative analysis with codes)	university students (B2)	China	English
Schierholz, S.J. (2015)	descriptive	—	Germany	—
Li, L. and H. Xu (2015)	mixed methods (n=32)	university students (B1)	China	English
Claassen, D. (2015)	qualitative (n=2 dictionaries)	—	Republic of South Africa	Afrikaans
Pujol, D. and M. Corrius (2013)	experiment (n=100)	university students (B1)	Spain	English
Lew, R., M. Grzelak and M. Leszkowicz (2013)	experiment (n=10)	higher education (5 participants: B2–C1; 5 participants: A2–B1)	Poland	English
Van der Merwe, M.F. (2012)	experiment (200)	native	Republic of South Africa	Afrikaans home language and Afrikaans first additional language
Chan, A.Y. (2012)	mixed methods (n=31)	university students (C1)	China	English

Boonmoh, A. (2012)	mixed methods (n=13) (think-aloud protocol, observation, and retrospective interviews)	university students (B1–B2)	Thailand	English
Beyer, H L. and A.K. Faul (2010)	quantitative (n=509)	first year students in higher education	Namibia	English
Boonmoh, A. (2010)	mixed methods	university students (B1–B2)	Thailand	English
De Schryver, G.-M. (2009)	descriptive investigating trends in <i>Lexikos</i> (qualitative)	—	—	—
De Schryver, G.-M., D. Joffe, P. Joffe and S. Hillewaert (2006)	log life analysis	—	Republic of South Africa	Swahili/English
Ekwa Ebanéga, G.-M. and F.T. Moussavou (2008)	quantitative (n=100)	native	Republic of South Africa	Gabonese
Al-Ajmi, H. (2008)	experiment (n=54)	university students (B2–C1)	Kuwait	English
Hadebe, S. (2004)	descriptive	—	Zimbabwe	Ndebele
Mdee, J.S. (1997)	qualitative (n=56) (self-reflections of participants)	university students (B1–C1)	Tanzania/ United Kingdom	German/French
Tono, Y. (1992)	experiment (n=239) (test)	secondary school students, university students (B1–B2)	Japan	English, Japanese

Publikasieaankondigings / Publication Announcements

Edward Finegan and Michael Adams. *The Cambridge Handbook of the Dictionary*. 2024, 801 pp. ISBN 978-1-108-83624-1 (Hardback), ISBN 978-1-108-81887-2 (Paperback), ISBN 978-1-108-86443-5 (eBook). Cambridge/New York, NY: Cambridge University Press. Price: US\$ 170.00 (Hardback), US\$ 170.00 (eBook). (Review in *Lexikos* 35(1).)

Pedro A. Fuertes-Olivera. *A Guide to Practical Online Lexicography*. 2025, 334 pp. ISBN: 978-1032731544 (Hardback), 9781003426998 (eBook). London/New York: Routledge. Price: US\$ 180.00 (Hardback).

Yongwei Gao. *A Dictionary of Contemporary English Neologisms and Trendy Words*. 2024, 406 pp. ISBN: 978-7-5213-5810-0 (Paperback). Beijing: Foreign Language Teaching and Research Press. Price \$9.32. (Review in *Lexikos* 35(1).)

Laura Giacomini and Valentina Piunno (Editors). *Patterns of Meaning in Lexicography and Lexicology*. 2025, XI + 319 pp. ISBN 978-3-11-148075-6 (Hardback), ISBN 978-3-11-154594-3 (eBook), ISSN 0175-9264. Lexicographica Series Maior 167. Berlin/Boston: Walter de Gruyter. Price: €109.95 (Hardback, eBook).

Theresa Kruse (Herausgeber). *Ein fachsprachliches elektronisches Lernwörterbuch. Konzeption: Prototypenstellung und Evaluation am Beispiel der Mathematik*. 2025, XIX + 356 pp. ISBN 978-3-11-138949-3 (Hardback), ISBN 978-3-11-159135-3 (eBook), ISSN 0175-9264. Lexicographica Series Maior 168. Berlin/Boston: Walter de Gruyter. Preis: €119.95 (Hardback, eBook).

Du Ruiqing and Jiang Yajun. *New Century Chinese–English Dictionary*. Third Edition. 2025, XXXVIII + 2490 pp. ISBN 978-7-5213-5418-8 (Hardback). Beijing: Foreign Language Teaching and Research Press. Price \$54.43. (Review in *Lexikos* 35(1).)

Stephen Turton. *Before the Word was Queer: Sexuality and the English Dictionary, 1600–1930*. 2024, xv + 333 pages. ISBN 978-1-316-51873-1 (Hardback). Cambridge: Cambridge University Press. Price: US\$ 94.67 (Hardback). (Review in *Lexikos* 35(1).)

VOORSKRIFTE AAN SKRYWERS

(Tree asseblief met ons in verbinding (lexikos@sun.ac.za) vir 'n uitvoeriger weergawe van hierdie instruksies of besoek ons webblad: <http://lexikos.journals.ac.za/>)

A. REDAKSIONELE BELEID

1. Aard en inhoud van artikels

Artikels kan handel oor die suiwer leksikografie of oor implikasies wat aanverwante terreine, bv. linguistiek, algemene taalwetenskap, terminologie, rekenaarwetenskap en bestuurskunde vir die leksikografie het.

Bydraes kan onder een van die volgende rubrieke geklassifiseer word:

(1) **Artikels:** Grondige oorspronklike wetenskaplike navorsing wat gedoen en die resultate wat verkry is, of bestaande navorsingsresultate en ander feite wat op 'n oorspronklike wyse oorsigtelik, interpreterend, vergelykend of krities evalueerend aangebied word.

(2) **Resensieartikels:** Navorsingsartikels wat in die vorm van 'n kritiese resensie van een of meer gepubliseerde wetenskaplike bronne aangebied word.

Bydraes in kategorieë (1) en (2) word aan streng anonieme keuring deur onafhanklike akademiese vakgenote onderwerp ten einde die internasionale navorsingsgehalte daarvan te verseker.

(3) **Resensies:** 'n Ontleding en kritiese evaluering van gepubliseerde wetenskaplike bronne en produkte, soos boeke en rekenaarprogramme.

(4) **Projekte:** Besprekings van leksikografiese projekte.

(5) **Leksikonotas:** Enige artikel wat praktykgerigte inligting, voorstelle, probleme, vrae, kommentaar en oplossings betreffende die leksikografie bevat.

(6) **Leksikovaria:** Enigeen van 'n groot verskeidenheid artikels, aankondigings en nuusvrystellings van leksikografiese verenigings wat veral vir die praktiserende leksikograaf van waarde sal wees.

(7) **Ander:** Van tyd tot tyd kan ander rubrieke deur die redaksie ingevoeg word, soos Leksikoprogrammatuur, Leksiko-opname, Leksikobibliografie, Leksikonuus, Lexikofokus, Leksiko-eerbewys, Leksikohuldeblyk, Verslae van konferensies en werksessies.

Bydraes in kategorieë (3)-(7) moet almal aan die eise van akademiese geskrifte voldoen en word met die oog hierop deur die redaksie gekeur.

2. Wetenskaplike standaard en keuringsprosedure

Lexikos is deur die Departement van Hoër Onderwys van die Suid-Afrikaanse Regering as 'n gesubsidieerde, d.w.s. inkomstegenererende navorsingstydskrif goedgekeur. Dit verskyn ook op die *Institute of Science Index (ISI)*.

Artikels sal op grond van die volgende aspekte beoordeel word: taal en styl; saaklikheid en verstaanbaarheid; probleemstelling, beredenering en gevolgtrekking; verwysing na die belangrikste en jongste literatuur; wesenlike bydrae tot die spesifieke vakgebied.

Manuskripte word vir publikasie oorweeg met dien verstande dat die redaksie die reg voorbehou om veranderinge aan te bring om die styl en aanbieding in ooreenstemming met die redaksionele beleid te bring. Outeurs moet toesien dat hulle bydraes taalkundig en stilisties geredigeer word voordat dit ingelewer word.

3. Taal van bydraes

Afrikaans, Duits, Engels, Frans of Nederlands.

4. Kopiereg

Nóg die Buro van die WAT nóg die African Association for Lexicography (AFRILEX) aanvaar enige aanspreeklikheid vir eise wat uit meewerkende skrywers se gebruik van materiaal uit ander bronne mag spruit.

Outeursreg op alle materiaal wat in *Lexikos* gepubliseer is, berus by die Direksie van die Woordeboek van die Afrikaanse Taal. Dit staan skrywers egter vry om hulle materiaal elders te gebruik mits *Lexikos* (AFRILEX-reeks) erken word as die oorspronklike publikasiebron.

5. Oorspronklikheid

Slegs oorspronklike werk sal vir opname oorweeg word. Skrywers dra die volle verantwoordelikheid vir die oorspronklikheid en feitelike inhoud van hulle publikasies. Indien van toepassing, moet besonderhede van die oorsprong van die artikel (byvoorbeeld 'n referaat by 'n kongres) verskaf word.

6. Gratis oordrukke en eksemplare

Lexikos is sedert volume 28 slegs elektronies beskikbaar op <http://lexikos.journals.ac.za>. Geen oordrukke of eksemplare is dus beskikbaar nie.

7. Uitnodiging en redaksionele adres

Alle belangstellende skrywers is welkom om bydraes vir opname in *Lexikos* te lewer en verkieslik in elektroniese formaat aan die volgende adres te stuur: lexikos@sun.ac.za, of Die Redakteur: LEXIKOS, Buro van die WAT, Postbus 245, 7599 STELLENBOSCH, Republiek van Suid-Afrika.

B. VOORBEREIDING VAN MANUSKRIP

Die manuskrip van artikels moet aan die volgende redaksionele vereistes voldoen:

1. Lengte en formaat van artikels

Manuskrip moet verkieslik in elektroniese formaat per e-pos of op rekenaarskyf voorgelê word in sagteware wat versoenbaar is met MS Word. Die lettersoort moet verkieslik 10-punt Palatino of Times Roman wees. Bydraes moet verkieslik nie 8 000 woorde oorskry nie.

Elke artikel moet voorsien wees van 'n opsomming van ongeveer 200 woorde en ongeveer 10 sleutelwoorde in die taal waarin dit geskryf is, sowel as 'n opsomming en sleutelwoorde in Engels. Engelse artikels van Suid-Afrikaanse oorsprong moet 'n opsomming en sleutelwoorde in Afrikaans hê, terwyl Engelse artikels van buitelandse oorsprong 'n tweede opsomming en sleutelwoorde in enigeen van die aangeduide tale mag gee. As die outeur dit nie doen nie, sal die redaksie 'n Afrikaanse vertaling voorsien. Maak seker dat die opsomming in die tweede taal ook 'n vertaling van die oorspronklike titel bevat.

2. Grafika

Figure, soos tabelle, grafieke, diagramme en illustrasies, moet in 'n gepaste grootte wees dat dit versoen kan word met die bladspieël van *Lexikos*, naamlik 18 cm hoog by 12 cm breed. Die plasing van grafika binne die teks moet duidelik aangedui word. Indien skryftekens of grafika probleme oplewer, mag 'n uitdruk van die manuskrip of 'n e-pos in .pdf-formaat aangevra word.

3. Bibliografiese gegewens en verwysings binne die teks

Kyk na onlangse nommers van *Lexikos* vir meer inligting. Buiten in spesiale gevalle moet verwysings na *Lexikos*-artikels tot twee of drie per artikel beperk word. Uitsonderings moet met die redakteur van *Lexikos* uitgeklaar word. Dit word gedoen om die status van *Lexikos* in verskeie internasionale indekse te behou.

4. Aantekeninge/voetnote/eindnote

Aantekeninge moet deurlopend in die vorm van boskrifte genommer en aan die einde van die manuskrip onder die opskrif **Eindnote** gelys word.

INSTRUCTIONS TO AUTHORS

(For a more detailed version of these instructions, please contact us (lexikos@sun.ac.za) or refer to our website: <http://lexikos.journals.ac.za/>)

A. EDITORIAL POLICY

1. Type and content of articles

Articles may treat pure lexicography or the implications that related fields such as linguistics, general linguistics, terminology, computer science and management have for lexicography.

Contributions may be classified in any one of the following categories:

- (1) **Articles:** Fundamentally original scientific research done and the results obtained, or existing research results and other facts reflected in an original, synoptic, interpretative, comparative or critically evaluative manner.
- (2) **Review articles:** Research articles presented in the form of a critical review of one or more published scientific sources. Contributions in categories (1) and (2) are subjected to strict anonymous evaluation by independent academic peers in order to ensure the international research quality thereof.
- (3) **Reviews:** An analysis and critical evaluation of published scientific sources and products, such as books and computer software.
- (4) **Projects:** Discussions of lexicographical projects.
- (5) **Lexiconotes:** Any article containing practice-oriented information, suggestions, problems, questions, commentary and solutions regarding lexicography.
- (6) **Lexicovaria:** Any of a large variety of articles containing announcements and press releases by lexicographic societies which are of particular value to the practising lexicographer.
- (7) **Other:** From time to time other categories may be inserted by the editors, such as Lexicosoftware, Lexicosurvey, Lexicobibliography, Lexiconews, Lexicofocus, Lexicohonour, Lexicotribute, Reports on conferences and workshops.

Contributions in categories (3)-(7) must all meet the requirements of academic writing and are evaluated by the editors with this in mind.

2. Academic standard and evaluation procedure

The Department of Higher Education of the South African Government has approved *Lexikos* as a subsidized, i.e. income-generating research journal. It is also included in the *Institute of Science Index (ISI)*.

Articles will be evaluated on the following aspects: language and style; conciseness and comprehensibility; problem formulation, reasoning and conclusion; references to the most important and most recent literature; substantial contribution to the specific discipline.

Manuscripts are considered for publication on the understanding that the editors reserve the right to effect changes to the style and presentation in conformance with editorial policy. Authors are responsible for the linguistic and stylistic editing of their contributions prior their submission.

3. Language of contributions

Afrikaans, Dutch, English, French or German.

4. Copyright

Neither the Bureau of the WAT nor the African Association for Lexicography (AFRILEX) accepts any responsibility for claims which may arise from contributing authors' use of material from other sources.

Copyright of all material published in *Lexikos* will be vested in the Board of Directors of the Woordeboek van die Afrikaanse Taal. Authors are free, however, to use their

material elsewhere provided that *Lexikos* (AFRILEX Series) is acknowledged as the original publication source.

5. Originality

Only original contributions will be considered for publication. Authors bear full responsibility for the originality and factual content of their contributions. If applicable, details about the origin of the article (e.g. paper read at a conference) should be supplied.

6. Free offprints and copies

Lexikos is only available electronically on <http://lexikos.journals.ac.za> from volume 28 onward. No offprints or copies are available.

7. Invitation and editorial address

All interested authors are invited to submit contributions, preferably in electronic format, for publication in *Lexikos* to: lexikos@sun.ac.za, or

The Editor: LEXIKOS
Bureau of the WAT
P.O. Box 245
7599 STELLENBOSCH, Republic of South Africa

B. PREPARATION OF MANUSCRIPTS

Manuscripts of articles must meet the following editorial requirements:

1. Format and length of articles

Manuscript should preferably be submitted in electronic format by email or on a disk, in software compatible with MS Word. The typeface used should preferably be 10-point Palatino or Times Roman. Contributions should not exceed **8 000 words**.

Each article must be accompanied by **abstracts** of approximately 200 words and approximately 10 **keywords** in the language in which it is written, as well as in **English**. English articles of South African origin should carry an abstract and keywords in Afrikaans, whilst English articles of foreign origin should carry a second abstract and keywords in any of the other languages mentioned. In cases where this is not done, the editors will provide an Afrikaans version. Ensure that the abstract in the second language also contains a **translation of the original title**.

2. Graphics

Figures such as tables, graphs, diagrams and illustrations should be in an appropriate size to be well accommodated within the page size of *Lexikos*, namely 18 cm high by 12 cm wide. The locations of figures within the text must be clearly indicated. If orthographic marks or graphics used in the text prove problematic, a printout of the manuscript or an email in .pdf format may be requested.

3. Bibliographical details and references in the text

Examine recent issues of *Lexikos* for details. Self-references to *Lexikos* should be limited to two or three per article, except in exceptional circumstances. Exceptions should be cleared with the editor of *Lexikos*. This is done to preserve the status of *Lexikos* in various international indices.

4. Notes/footnotes/endnotes

Notes must be numbered consecutively by superscript numbers and grouped together at the end of the manuscript under the heading **Endnotes**.