http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)

Lexikos 34

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)

Lexikos 34

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

AFRILEX African Association for Lexicography

AFRILEX-REEKS 34:2024
AFRILEX SERIES 34:2024

BURO VAN DIE WAT
STELLENBOSCH

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)

Uitgewer Publisher

BURO VAN DIE WAT
Posbus 245
7599 STELLENBOSCH

Kopiereg © 2024 deur die uitgewer
Alle regte streng voorbehou
Eerste uitgawe 2024

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

ISBN 978-1-990998-53-9
ISSN 2224-0039

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

Licenced 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 x
Foreword xiii
André H. du Plessis

'n Woord van AFRILEX xvi
A Few Words from AFRILEX xvii
Langa Khumalo

Redaksionele doelstellings xviii
Editorial Objectives xix

Artikels / Articles

Der Effizienz- und Intelligenzbegriff in der Lexikographie und künstlichen Intelligenz: kann ChatGPT die lexikographische Textsorte nachbilden? 51
Iván Arias-Arias, María José Domínguez Vázquez und Carlos Valcárcel Riveiro

Exploring Dictionary Preferences: A Comparative Study of EFL and GFL Learners in Hungarian Higher Education 419
Balázs Fajt

This, *Thing*, *Fervor*, *Fulfilment*: The Treatment of Pronunciation and Spelling in Dictionaries of the Slovenian Immigration 141
Donna M.T.Cr. Farina, Marjeta Vrbinc and Alenka Vrbinc

Making Lexicography Sustainable: Using ChatGPT and Reusing Data for Lexicographic Purposes 123
Pedro A. Fuertes-Olivera

Algorithmic Complexity and Learnability in German Monolingual Learner Lexicography. A Case Study 166
Alberto Galván-Santana

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)

How Can We Raise Strategic Dictionary Use in the Classroom: The Effect of a Dictionary Awareness Program on Dictionary Use Strategies 99
Zoe Gavriilidou, Angelos Markos and Evanthia Konstantinidou

Aspekte van inligtingsonttrekkingstrukture in aanlyn woordeboeke 355
Rufus H. Gouws en Theo J.D. Bothma

The Inclusion of Neologisms in the Revision of the *Grand Dictionnaire Chinois–Français Contemporain* 469
Fang Huang and Jianhua Huang

Training an AI-based Writing Assistant for Spanish Learners: The Usefulness of Chatbots and the Indispensability of Human-assisted Intelligence 21
Ángel Huete-García and Sven Tarp

Adaptation and Validation of the Strategy Inventory for Electronic Dictionary Use (S.I.E.D.U.) for Chinese Learners 245
Lingling Li, Hui Wang and Hai Xu

Using Generative AI to Provide High-Quality Lexicographic Assistance to Chinese Learners of English 397
Qian Li and Sven Tarp

Using Semi-automated Term Extraction for IsiNdebele Health Terminology 269
Nomsebenzi Malele and Sonja Bosch

Bridging across Polysemic Senses in Bilingual Specialized Dictionaries for ESP Learners 486
Huaguo Lu and Yundong Geng

Grammatical Data in the *Dictionary of Montenegrin National and Literary Language* 218
Sonja Nenezić

Dictionaries in Context, Context in Dictionaries: Legal Translation Tools 288
Sandro Nielsen

Making African Dictionaries More African 309
Thapelo J. Otlogetswe

Academic Word Families in Online English Dictionaries 437
Geraint Paul Rees

Le Processus de Numérisation de la Lexicographie en Roumanie: Présent et Perspectives 1
Elena Isabelle Tamba

L'Evolution de la Terminologie de la Plasturgie entre 1963–2018: Analyse Diachronique et Synchronique 371
Valentina-Nicoleta Văsioiu et Marilena Milcu

Dictionary Use Training in Secondary School EFL Textbooks in Taiwan 77
Wai-on Law

Projekte / Projects

The Current State of the OBI DICT Project: A Bilingual e-Dictionary of Oracle-Bone Inscriptions with AI Image Recognition 198
Yang Jin and Shuo Wen

Leksikonotas / Lexiconotes

Defining Feminine Personal Nouns in Polish: A Practical and Postulative Overview Based on the *Dictionary of Polish Female Nouns* 236
Agnieszka Małocha

On the Inclusion of Neologisms in *Oxford Advanced Learner's Dictionary* (10th edition) 41
Anmin Wang and Xi Chen

Leksikovaria / Lexicovaria

A Scoping Review of Studies into Dictionary Use and Language Learning 331
Xiaoshuai Ge, Songshan Zhang, Hai Xu and Xian Zhang

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)

Resensies / Reviews

Hou Min. *A Dictionary of Chinese Neologisms (2000–2020)* 189
Yongwei Gao

Heming Yong, Jing Peng and Xiangming Zhang. *Chinese Lexicography in the Twentieth Century* 213
Yuanwen Zhang

Publikasieaankondigings / Publication Announcements 509

Voorskrifte aan Skrywers 510
Instructions to Authors 511

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)

Voorwoord

Soos 2023, het 2024 opwindende veranderinge en interessante uitdagings ingehou. Dit was 'n bedrywige jaar vir die *Lexikos*-span en 'n lywige uitgawe het die lig gesien. My tweede jaar as hoofredakteur het vir leergeleenthede sowel as oomblikke van genot gesorg. Om op hierdie unieke manier te kan verseker dat die leksikografiese diskoers gestimuleer word, is 'n voorreg en lekkerte wat nie elke leksikograaf noodwendig ervaar nie. Dit bly 'n eer om aan die stuur van 'n tydskrif met 'n wesenlike internasionale aansien te wees. Ek wil graag my opregte waardering aan die AFRILEX-raad en my mederedakteurs oordra vir die vertroue wat hulle in my stel, asook vir hulle deurlopende ondersteuning. Baie dankie ook aan AFRILEX vir die onderstand wat die assosiasie aan *Lexikos* verleen.

Lexikos is, soos telkens in die verlede vasgestel, 'n spanpoging en dit is regtig aangenaam om deel van so 'n uiters bekwame span te wees. 'n Spesiale woord van dank aan proff. Elsabé Taljard en Dion Nkomo vir die uitstekende werk wat hulle as mederedakteurs gedoen het, asook die waardevolle raad en bystand wat hulle gebied het. Baie dankie ook aan dr. Steve Ndinga-Koumba-Binza vir sy volgehoue en knap werk as Franse redakteur, ten spyte van nog 'n jaar met verskeie swarighede en struikelblokke. Dankie ook aan dr. Dané Claassen vir die goeie werk met die resensies. Resensies is 'n integrale deel van *Lexikos* se aanbieding en haar moeite daarmee gaan nie ongesiens verby nie. 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 uitgawe, die administratiewe bystand, asook vir hulle bekwaamheid met al die ad hoc-take wat deur die loop van die jaar opgeduik het. Dit is ook hier waar ek moet noem dat die *Lexikos*-span vanjaar twee nuwe lede bygekry het, te wete proff. Sonja Bosch en Michele van der Merwe. Prof. Bosch sluit by ons aan as Duitse redakteur en prof. Van der Merwe as mederedakteur. Dit is vir *Lexikos* 'n besondere voorreg om nog twee wêreldklas akademici en redakteurs in die redaksie te verwelkom. Beide proff. Bosch en Van der Merwe bring jare se leksikografiese en akademiese ervaring en hulle aanstelling is weliswaar 'n aanwins vir die span. Namens *Lexikos* en AFRILEX wil ek hulle hartlik bedank vir hulle bereidwilligheid en toewyding en vir die uitnemende werk wat hulle vanjaar gedoen het.

Die aanstelling van proff. Bosch en Van der Merwe het dit ook vir die redaksie moontlik gemaak om die redaksionele werksverdeling en struktuur te hersien. Daar is besluit dat ek vir die afsienbare toekoms die posisie van hoofredakteur sal beklee, met sowel redaksionele as oorhoofse administratiewe pligte. Proff. Bosch, Nkomo, Taljard en Van der Merwe en dr. Ndinga-Koumba-Binza dien voorts as mederedakteurs, en dr. Claassen as resensieredakteur. Hierdie herstrukturering is ter voordeel van die redaksionele werksverdeling, aangesien die mederedakteurs op hulle nodige akademiese en redaksionele take kan fokus met aan-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)

sienlik minder van die administratiewe las, terwyl ek die administratiewe sake saam met mee. Harteveld en Van der Westhuizen kan bestuur. Ek sal steeds my verwagte pligte as redakteur nakom, maar met groter ondersteuning van die mederedakteurs. Hierdie struktuur het ook tot gevolg dat die stelsel van roterende hoofredakteur tot 'n einde kom, wat vir kontinuïteit en verbeterde koördinasie tussen die hoofredakteur en mederedakteurs sorg. Tot dusver het hierdie struktuur goed gewerk en die hoop is dat dit 'n stabiele en volhoubare redaksionele en administratiewe grondslag vir die tydskrif sal verseker.

'n Tweede groot verandering wat van vanjaar plaasgevind het, is die hernuwing van *Lexikos* se adviesraad en redaksiekomitee. Die adviesraad en redaksiekomitee is veral steunpilare vir die redaksie, maar moet van tyd tot tyd hersien word om te verseker dat *Lexikos* aan die wetenskaplike en administratiewe vereistes van o.a. die Suid-Afrikaanse Departement van Hoër Onderwys en Opleiding voldoen. Hiermee word ook verseker dat die lede van hierdie liggame steeds aktief in die vakgebied is. Ons is die uitgaande adviesraad en redaksiekomitee baie dank en waardering verskuldig vir hulle jarelange diens en ondersteuning. Namens die redaksie wil ek ook graag die inkomende raad en komitee verwelkom en by voorbaat bedank vir hulle gewilligheid, steun en goedgesindheid.

Lexikos sou natuurlik nie sonder sy outeurs en hulle bydraes kon voortbestaan nie. Vanjaar se uitgawe bevat interessante, unieke en relevante bydraes wat heelwat stof tot nadenke bied. Die gebruik van of rol van KI in die leksikografie is veral tans 'n belangrike onderwerp wat ook in hierdie uitgawe te voorskyn kom. Daar is vier artikels wat hierdie onderwerp op insiggewende maniere en vanuit verskillende perspektiewe bespreek. Die artikels wat Afrika-leksikografie sterk bevorder, lewer eweneens waardevolle bydraes. Verder is daar ook betekenisvolle bydraes oor pedagogiese leksikografie, aanlyn woordeboeke, historiese woordeboeke, terminologie, gespesialiseerde leksikografie, en sosiale en unieke kwessies in die metaleksikografie en leksikografiepraktyk. Naas ons outeurs uit Afrika, is daar 'n wonderlike verskeidenheid stemme uit verskillende dele van die wêreld, soos China, Denemarke, Griekeland, Roemenië, Spanje en Slowenië, wat ook die omvang en diversiteit van ons vakgebied illustreer. Baie dankie aan die outeurs wat *Lexikos* gekies het om hulle navorsing en kundigheid te deel. Die redaksie is ook dankbaar vir al die outeurs se samewerking en positiewe gesindhede.

Lexikos is in besonder van sy kranige portuurbeoordelaars afhanklik. Die redaksie is innig dankbaar vir die onskatbare bydraes van elke keurder wat vanjaar by portuurbeoordeling betrokke was, en daardeur verseker het dat die hoë wetenskaplike standaard van hierdie tydskrif gehandhaaf word.

Ons het vanjaar met groot leedwese verneem van die heengaan van twee leksikografiese reuse: Patrick Hanks, wat in Februarie oorlede is, en Reinhard Hartmann, in September. Die baanbrekerswerk wat hulle in die leksikografie sowel as linguistiek gedoen het, kan nie onderskat word nie. Beide Hanks en Hartmann het ongetwyfeld leksikografiese praktyke, teorieë, navorsing en aktiwiteite op verskeie vlakke sinvol en beduidend verander, verbeter en uitgedaag. Hulle toewyding aan die leksikografie, hulle ongelooflike passie en betrokken-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)

heid by verskeie aspekte daarvan, moet as dié voorbeeld vir enige aspirantleksikograaf dien. Dit is 'n enorme verlies vir ons vakgebied en ons wil graag ons meegevoel aan elkeen van hulle geliefdes, vriende, kollegas en medewerkers oordra.

Die *Lexikos*-span het vanjaar 'n nuwe protokol vir die publisering van spesiale uitgawes ingestel, die redaksionele beleid verfyn en 'n nuwe aanlyn verwysingsgids vir outeurs beskikbaar gestel. Hierdie klein, dog belangrike verbeteringe, maak dat *Lexikos* op alle vlakke op standaard bly en dat hierdie standaarde in die toekoms gehandhaaf kan word. Mag hierdie uitgawe 'n weerspieëling hiervan wees en ook 'n sinvolle bydrae tot die leksikografie as vakgebied lewer.

André H. du Plessis
Redakteur

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)

Foreword

Like 2023, 2024 brought exciting changes and interesting challenges. It was a productive year for the *Lexikos* team which saw a sizeable volume published. My second year as editor-in-chief provided opportunities to learn as well as moments of enjoyment. Being in a position to stimulate the lexicographic discourse in this unique way is a privilege and a pleasure that does not befall many a lexicographer. It is an enduring honour to be at the helm of a journal with a substantial international standing. I would like to express my heartfelt appreciation to the AFRILEX board and my associate editors for their continued support and trust in me. A big thank you to AFRILEX for the financial aid also provided to *Lexikos*.

Lexikos is, as determined time and time again, a team effort and it is especially gratifying to be part of such a highly proficient team. A special word of thanks to Profs Elsabé Taljard and Dion Nkomo for the outstanding work delivered as associate editors, as well as for all their valuable guidance and assistance. Thank you also to Dr Steve Ndinga-Koumba-Binza for his untiring and deft work, despite another year of personal difficulties and obstacles. Thank you to Dr Dané Claassen for the good work with the reviews. Reviews are an integral part of *Lexikos'* offering and Dr Claassen's efforts do not go unnoticed. On behalf 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 issue, 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. It is also here that I have to note two additions to the *Lexikos* team. Prof Sonja Bosch joins the team as German editor and Prof Michele van der Merwe as associate editor. It is a distinct honour for *Lexikos* to welcome two more world-class academics and editors to the editorial team. Profs Bosch and Van der Merwe bring years of lexicographic and academic experience, and they are truly valuable additions to the team. On behalf of *Lexikos* and AFRILEX, I would like to thank them for their willingness and devotion to the journal, and for the excellent work done.

The appointments of Profs Bosch and Van der Merwe enabled the editorial team to review the editorial structure and subsequent division of work. The decision was made that I will continue as editor-in-chief for the foreseeable future, with both editorial and administrative responsibilities. Profs Bosch, Nkomo, Taljard and Van der Merwe, and Dr Ndinga-Koumba-Binza will continue to serve as associate editors, while Dr Claassen remains reviews editor. This restructuring is advantageous for the division of work as associate editors can now focus on their academic and editorial responsibilities without bearing a heavy administrative load as well. I, along with Mss Harteveld and Van der

Westhuizen, will manage much of the administrative matters. I will of course still fulfil my responsibilities as editor, but with more support from the other editors. This new structure also means that the system of rotating editorship comes to an end. This will guarantee more continuity and better coordination between editors. Thus far the new structure has worked well, and we are hopeful that it will lay a stable and sustainable editorial and administrative foundation for the journal.

Another major change that occurred this year is the renewal of *Lexikos'* advisory board and editorial committee. The advisory board and editorial committee are mainstays for the editorial team, but these bodies need to be reviewed and renewed from time to time to ensure that the journal meets the necessary academic and administrative requirements determined by, inter alia, the South African Department of Higher Education and Training. This renewal process is also undertaken to ensure that the members of these bodies are still active in the field of lexicography. We owe a great debt of gratitude and appreciation to the outgoing advisory board and editorial committee for their years of service and support. On behalf of the editorial team, I would like to welcome the incoming members of these bodies and thank them in advance for their readiness, assistance and goodwill.

Lexikos would of course not be able to exist without its authors and their contributions. This year's volume contains interesting, unique, relevant and thought-provoking contributions. The use and role of AI in lexicography is an incredibly significant and important topic that also appear in this volume. It is insightfully tackled in four articles with each presenting a distinct perspective. The articles that strongly promote African lexicography, make valuable contributions. Furthermore, there are articles of note that discuss pedagogical lexicography, online dictionaries, historical dictionaries, terminology, specialised lexicography, and social and unique issues in both theoretical and practical lexicography. The local contributions are joined by a wonderful array of voices from various parts of the globe, such as China, Denmark, Greece, Romania, Spain and Slovenia, which typify the scope and diversity of our field. A big thank-you to the authors who chose *Lexikos* as a platform to share their expertise and research. The editors are grateful for all the authors' worthy contributions and for their positive cooperation and attitudes.

Lexikos is dependent on its expert peer reviewers. The editors are 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 two lexicographic giants: Patrick Hanks, who passed away in February, and Reinhard Hartmann, who passed in September. Their pioneering work in the fields of lexicography and linguistics cannot be underestimated. Undeniably, both Hanks and Hartmann notably and meaningfully transformed, improved and challenged lexicographic practices, theories, research and activities on many different levels. Their devotion to, passion for and their involvement in the many distinct aspects of lexicography, can serve as the example for any aspiring lexicographer. It is an enormous loss for our discipline, and we send our condolences to all their loved ones, friends, colleagues and collaborators.

Lastly, this year the *Lexikos* team implemented a new protocol for the publication of special issues, refined the editorial policy and made a new online reference guide available to authors. These small, yet valuable improvements, ensure that *Lexikos'* standards are maintained and that they will be upheld in the future. May this year's volume be a worthy contribution to the field of lexicography, as well as a reflection of the aforementioned high standards.

André H. du Plessis
Editor

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)

'n Woord van AFRILEX

Die *African Association for Lexicography* (AFRILEX) se Raad wens die *Lexikos*-span, onder leiding van die hoofredakteur, André du Plessis, en die uitgewer, die Buro van die WAT, hartlik geluk met die suksesvolle publikasie van vanjaar se *Lexikos*. Die Raad is deurgaans bewus van die deeglike werk wat deur die redaksie gedoen word. Hulle 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. Ná die suksesvolle afhandeling van die 28ste Internasionale AFRILEX-kongres, waar die Universiteit van Pretoria die gasheer was en wat ook 'n hibridiese Globalex-werkswinkel ingesluit het, is daar 'n toename in *Lexikos*-bydraes van outeurs buite (Suid-)Afrika. Hoewel dit die internasionele omvang en strekking van die tydskrif vergroot, is daar ook 'n doelgerigte poging om by plaaslike navorsers belangstelling te kweek en bydraes aan te moedig.

Die 28ste Internasionale AFRILEX-kongres het ook twee spesiale sessies bevat: een vir die Nasionale Leksikografie-eenhede en een vir plaaslike woordeboekuitgewers. Hierdie twee sessies het vir 'n unieke samesmelting van woordeboekteorie, -praktyk, -beginsels en -gebruik in (Suid-)Afrika en elders gesorg. Gevolglik ontstaan die potensiaal vir 'n toenemende en interessante navorsingskoers wat aandag aan hierdie samesmelting kan gee, en waaruit vele sinvolle bydraes in die toekoms sal spruit.

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

The African Association for Lexicography (AFRILEX) Board wants to congratulate the *Lexikos* team, under guidance of the Editor-in-Chief, André du Plessis, and the publisher, the Bureau of the WAT, for the successful publication of this year's volume. The Board remains mindful of the sterling work that is done by the *Lexikos* editors. Their stringency, meticulous attention to detail and expertise in the field has helped maintain the very high standards of the journal.

Lexikos remains an important mouthpiece of the cutting-edge research that is done by the AFRILEX members in Africa and globally. Following the successful hosting of the 28th International AFRILEX Conference at the University of Pretoria, that also included the Globalex workshop as a special hybrid session, there is an increased number of articles published in this *Lexikos* volume from authors outside South Africa. While this increases the international footprint and reach of the journal, there is a concerted effort to cultivate and grow interest and contributions from (South) African researchers.

The 28th International AFRILEX Conference also saw two more special sessions. These were the National Lexicography Units' session and the Publishers' session. These two special sessions brought a unique conflation of theory, practice, principles, and usage of dictionaries in (South) Africa and globally. As a result, there is potential for an interesting and growing scholarship that looks at these intersections, and we look forward to more interesting publication outcomes in future.

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

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)

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 interdissiplinêre 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.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)

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.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1857 (Article)

Le Processus de Numérisation de la Lexicographie en Roumanie: Présent et Perspectives

Elena Isabelle Tamba, *Université Alexandru Ioan Cuza de Iasi, Roumanie / Institut de Philologie Roumaine Alexandru Philippide, Académie Roumaine, Filiale de Iasi, Roumanie (isabelle.tamba@gmail.com) (https://orcid.org/0000-0001-5372-2363)*

Résumé: En Roumanie, la recherche lexicographique est en développement continu, dans un processus naturel de numérisation. Cette étape est absolument obligatoire pour la création d'instruments et de ressources électroniques, nécessaires au soutien de la langue et de la culture roumaines.

Les spécialistes roumains en linguistique et informatique appliquée, ainsi que dans les domaines de la linguistique computationnelle, ont initié des projets de recherche par lesquels ils valorisent les ressources non numérisées en les acquérant sous forme électronique ou par lesquels ils créent des dictionnaires et de nouvelles ressources et instruments directement sous forme électronique.

Le processus de numérisation place la lexicographie académique roumaine à un niveau comparable à la lexicographie internationale et permet la connexion avec des projets lexicographiques de l'étranger et l'inclusion de la Roumanie dans la sphère d'intérêt des grands réseaux lexicographiques internationaux.

Mots-clés: LEXICOGRAPHIE ROUMAINE, E-LEXICOGRAPHIE, CORPUS LEXICOGRAPHIQUE ÉLECTRONIQUE, NUMÉRISATION, RESSOURCES LEXICOGRAPHIQUES NUMERISÉE, ÉVOLUTION, PERSPECTIVES

Abstract: Digitalization Process of the Lexicography in Romania: Present and Perspectives. In Romania, the lexicographic research is in a continuous development, in a natural process of digitalization. This step is absolutely mandatory for creating electronic instruments and resources, which are necessary for supporting the Romanian language and culture.

The Romanian academic specialists in linguistics and applied informatics, as well as in computational linguistics fields, have initiated research projects by which they valorise the non-digitized resources by acquiring them in electronic formats or by which they create dictionaries and new resources and instruments directly in electronic format.

The digitalization process put the Romanian academic lexicography at a level comparable to the international lexicography and allows the connection with lexicographic projects from abroad and the inclusion of Romania in the sphere of interest of the great international lexicographic networks.

Keywords: ROMANIAN LEXICOGRAPHY, E-LEXICOGRAPHY, ELECTRONIC LEXICOGRAPHIC CORPUS, DIGITALIZATION, DIGITAL LEXICOGRAPHICAL RESOURCES, EVOLUTION, PERSPECTIVES

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1857 (Article)

1. Introduction

La langue représente une expression de la culture, un facteur déterminant de l'identité et, en même temps, elle offre la possibilité de pouvoir communiquer. La lexicographie soutient la langue, en mettant dans les articles de dictionnaire l'histoire de chaque mot et, de cette manière, elle raconte l'histoire de la Langue. C'est pourquoi rendre plus facile l'accès à l'information est de plus en plus important. À travers l'informatisation des ressources lexicographiques, la langue roumaine s'inscrit dans cette direction de rendre l'information accessible aux personnes intéressées par la recherche de la langue roumaine / par l'étude du roumain.

La lexicographie mondiale connait un vaste processus de changement, de modernisation des moyens de rédaction, de consultation, etc., à travers des approches qui impliquent l'interconnexion de différents domaines de recherche. Hanks (2013) fait référence aux ordinateurs et aux corpus comme le deuxième facteur le plus influent, après l'invention de l'imprimerie, en lexicographie. C'est pourquoi la nouvelle technologie du langage et du corpus a fourni aux lexicographes et aux utilisateurs des outils innovants pour compiler et consulter des dictionnaires (cf. Abdelzaher 2022).

En Roumanie, surtout au cours de la dernière décennie, la recherche lexicographique (et linguistique, d'ailleurs) connait un développement continu, dans un processus naturel de numérisation. Cette étape est absolument obligatoire pour créer des instruments électroniques et des ressources nécessaires au soutien de la langue et de la culture roumaines.

Ce processus de numérisation de la recherche lexicographique / linguistique suppose plusieurs étapes:

- changement du format classique / en papier au format numérique, ce qui conduit à la création de corpus lexicographiques (ou linguistiques), par plusieurs approches de recherche
- numérisation des dictionnaires en papier
- création / développement des corpus lexicographiques
- création / développement des corpus linguistiques, annotés du point de vue morphologique, syntaxique et sémantique
- création des ressources lexicographiques directement en format numérique, par
- création des dictionnaires directement en format numérique
- utilisation des programmes de rédaction des dictionnaires
- utilisation des programmes pour extraire des citations, etc.

Cet article envisage la description des principaux projets et résultats dans le domaine de la lexicographie roumaine numérique, tant pour les dictionnaires contemporains que prémodernes.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1857 (Article)

2. La lexicographie numérique en Roumanie

Les professionnels roumains en lexicographie et en linguistique informatique ont initié des projets de recherche par lesquels ils veulent, d'une part, valoriser les ressources non-numérisées, en les acquérant dans des formats électroniques et, d'autre part, créer de nouvelles ressources et instruments lexicographiques / linguistiques directement en format numérique.

La plupart des efforts de numérisation des recherches lexicographiques ont été faits sous les auspices des Instituts de l'Académie Roumaine[1]; mais, plus récemment, les centres de recherche de certaines universités du pays s'y sont impliqués[2].

À présent, en Roumanie il y a de divers projets lexicographiques (ou linguistiques) numérisés, réalisés à partir des initiatives académiques — on y inclut non seulement des instituts de langue et de littérature roumaine et les instituts d'informatique de l'Académie Roumaine, mais aussi des centres de recherche de certaines universités de Roumanie et des bibliothèques. Mais il y a aussi des initiatives privées (des projets initiés par de volontiers[3]; des projets de certaines maisons d'édition, etc.).

2.1 La numérisation du *Dicționarul limbii române* de l'Académie roumaine

Dicționarul limbii române [Le Dictionnaire de la langue roumaine] est l'ouvrage lexicographique le plus important pour le roumain, étant rédigé et édité par l'Académie Roumaine[4]. C'est pourquoi la création d'un format numérique, accessible aux scientifiques et à tous ceux qui s'intéressent à l'apprentissage ou à l'étude de la langue roumaine, en Roumanie ou à l'étranger, est devenue, dans la société numérique et multiculturelle, une étape absolument nécessaire à franchir.

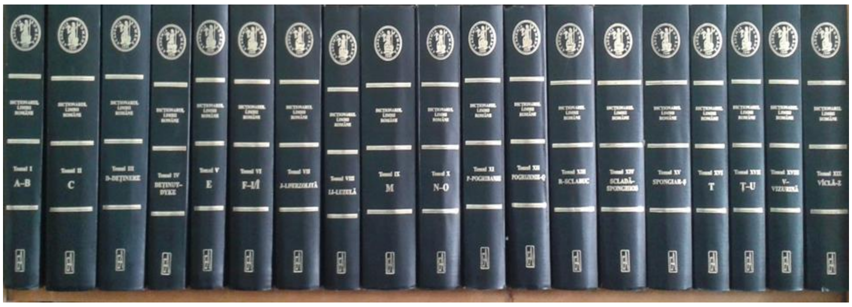


Figure 1: *Dicționarul limbii române*. Édition anastatique, 2010

Les grandes cultures disposent depuis de nombreuses années de dictionnaires et de corpus de textes au format numérique. Pour une meilleure compréhension des dimensions du *Dictionnaire-trésor de la langue roumaine*, on présente quelques-unes de ses données statistiques, comparées à d'autres grands dictionnaires européens:

— La compilation du *Dicționarul limbii române* a commencé il y a 115 ans. La première édition fut publiée en deux séries — DA (1913–1944) et DLR (1965–2010), en 14 tomes, 37 volumes, 20 000 pages de type lexique (ayant entre 7 000 et 11 000 caractères/page), plus de 175 000 entrées (avec variantes) et plus de 1 300 000 citations; l'élaboration du format numérique: première tentative, eDTLR, 2007–2010 (à présent, eDTLR est un projet repris, en plein déroulement); une autre tentative de numérisation, réussie cette fois-ci — dans le projet CLRE[5]. La deuxième édition, révisée et ajoutée, (en format numérique et en papier) est en cours de rédaction, directement en format numérique (DLR^2).

Afin d'établir une comparaison entre le dictionnaire-trésor roumain et d'autres dictionnaires importants, on présente quelques informations statistiques:

— *Dictionnaire de l'Académie Française* (DAF), 1694 — la première édition imprimée; 9 éditions, consultables en ligne, 55.000 entrées;

— *Diccionario de la lengua española de la Real Academia Española* (DRAE), 1780 — la première édition imprimée; la 23e édition — 2014; 93 111 lemmes; le premier format numérique: 1992;

— *Deutsches Wörterbuch der Grimm* (DWB), 1838–1961, 32 volumes, 350.000 entrées et variantes; le premier format numérique: 1997–2004;

— *Oxford English Dictionary* (OED), la première édition — 1928, 20 volumes (la deuxième édition — 1989), 301.100 entrées, 2.412.400 exemples; le premier format numérique: 1988;

— *Trésor de la Langue Française* (TLF), XIXe–XXe siècle, 1971-1994 — la première édition imprimée; 16 volumes, 100.000 entrées, 270.000 définitions, 430.000 exemples; le premier format numérique: 1990–2004;

— *Tesoro della lingua italiana delle origini* (TLIO) — en ligne, 44.000 entrées (37864 publiés en ligne de 57.000).

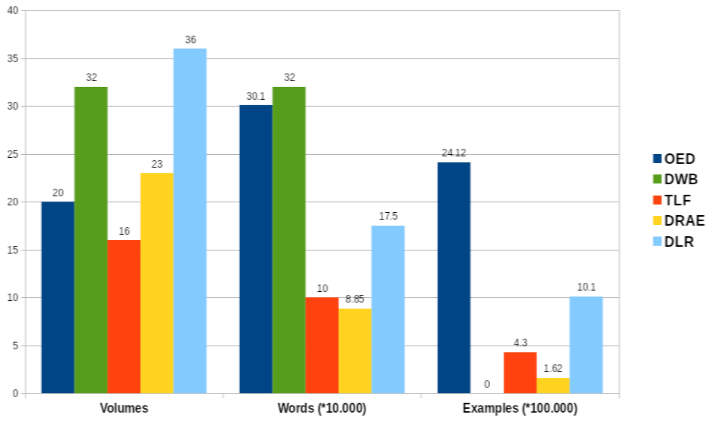


Figure 2: DLR vs. d'autres dictionnaires européens[6]

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1857 (Article)

À partir des données ci-dessus, on peut observer le fait que *Dicționarul limbii române* est similaire, en ce qui concerne la conception et la réalisation, aux autres dictionnaires européens et sa numérisation représente, ainsi, une étape normale dans l'évolution de la lexicographie roumaine.

Pour mieux comprendre le parcours qui a conduit à l'étape actuelle de la lexicographie dans le cadre de l'Académie Roumaine, on va mentionner, dans ce qui suit, les projets qui ont contribué à façonner une stratégie concernant la numérisation de la recherche lexicographique académique roumaine[7]:

— 2003–2005: le projet pilote de numérisation du DLR, *Dicționarul limbii române (DLR) în format electronic. Studii privind achiziționarea* [Le Dictionnaire de la langue roumaine (DLR) en format électronique. Études sur la création du format numérique][8];

— 2006–2007: première édition électronique d'un ancien texte roumain de la Bibliographie du DLR, *Monumenta linguae Dacoromanorum. Bible 1688. Pars VII. Regum I, Regum II* (https://biblia1688.solirom.ro/7/);

— 2007–2008: mise à jour numérisée des premiers éléments de la série DA selon les normes de rédaction DLR dans un projet nommé *DLRI. Bază lexicală informatizată. Derivate în -ime şi -işte* [DLRI. Base lexicale informatisée. Dérivés en –ime et –iște][9];

— 2007–2010: première tentative de numérisation complète du DLR dans le projet homonyme, l'objectif principal étant l'acquisition en format numérique de la variante intégrale du *Dictionnaire-trésor de la langue roumaine*, grâce à la rétro-numérisation;

— 2010: initiation de la documentation et rédaction directement sous format électronique de la deuxième édition du *Dicționarul limbii române. DLR^2.* [Le Dictionnaire de la langue roumaine. DLR^2] — projet fondamental et prioritaire de l'Académie Roumaine; la deuxième édition de ce dictionnaire a été initié en 2010 et elle se rédige entièrement en format numérique (https://dlri.ro/); en parallèle, une version en papier est également publiée;

— 2010–2013: premier projet dont l'objectif principal est la création d'un corpus lexicographique diachronique: *CLRE. Corpus lexicografic românesc esențial* [CLRE. Corpus lexicographique roumain essentiel];

— 2014: inclusion du projet CLRE dans le plan de recherche de l'Académie Roumaine. *CLRE. Corpus lexicografic românesc electronic* [Corpus lexicographique roumain électronique] — projet fondamental de l'Académie roumaine;

— 2019–présent: le passage d'ouvrages lexicographiques numérisés, publiés isolément, à des ouvrages publiés unitairement, grâce à l'utilisation des normes et d'outils de travail communs, sur la plate-forme SOLIROM[10], qui appartient aux instituts à profil philologique de l'Académie Roumaine et

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1857 (Article)

qui permet la collaboration entre spécialistes, en utilisant les outils et les ressources créés, ainsi que la publication intégrée des travaux développés sous les auspices de l'Académie Roumaine.

À présent, *Dicționarul limbii române* édité sous les auspices de l'Académie Roumaine connait trois variantes numériques, réalisées ou en train d'être réalisées dans trois projets différents, qui vont être connectées dans *CLRE. Corpus lexicografic românesc electronic*:

(a) **DLR²** — la rédaction de la deuxième édition de *DLR. Dicționarul limbii române. Ediția a doua revăzută și adăugită* [DLR. Dictionnaire de la langue roumaine. Deuxième édition révisée et ajoutée] a commencé en 2010 et est ou sera faite entièrement directement en format numérique, à l'aide d'une interface d'écriture XML, dans le vocabulaire TEI, et publiée en ligne, au fur et à mesure que les fascicules sont finalisés — https://dlri.ro/. Ce projet est réalisé dans les Départements de Lexicographie des instituts de langue et de littérature roumaine de l'Académie Roumaine. La parution du premier fascicule de la deuxième édition, révisée et complétée, du *Dictionnaire de la langue roumaine*. Volume I. Lettre A. Fascicule 1 (*A–Abzițui*), rédigé en format numérique par l'équipe de lexicographes de l'Institut de Linguistique «Iorgu Iordan — Al. Rosetti», de l'Académie Roumaine, à Bucarest (https://dlri.ro/) a été suivi par la parution du fascicule en format classique, sur papier.

(b) **DA, DLR** — rétro-numérisation, dans le projet CLRE (qui suppose scanner et traiter les informations dans la plate-forme, qui permet pour l'instant des consultations au niveau de l'entrée et au niveau de la catégorie lexico-grammaticale et l'affichage de l'image de la page / des pages du dictionnaire — voir les informations sur CLRE dans le sous-chapitre suivant).

(c) **eDTLR** — rétro-numérisation dans le projet ayant le même nom. La première tentative de numérisation complète du DLR a eu lieu dans le projet *Dicţionarul tezaur al limbii române în format electronic* (eDTLR) [Dictionnaire-trésor de la langue roumaine en format numérique (eDTLR)] (2007–2010), première version (partenariat entre les Instituts de l'Académie Roumaine et la Faculté d'Informatique de l'Université «Alexandru Ioan Cuza» à Iași) — l'objectif principal étant l'acquisition en format numérique de la variante intégrale du *Dictionnaire-trésor de la langue roumaine*, par des recherches qui suppose scanner, OCR-iser, corriger, analyser et introduire les données dans une plate-forme de consultation, qui permet des recherches complexes, dans tout le corps de chaque article lexicographique. À partir de 2021, le projet a été repris, dans l'Académie Roumaine — Branche Iași, en tant que projet prioritaire, afin de corriger et compléter la version numérique eDTLR et de l'inclure dans CLRE.

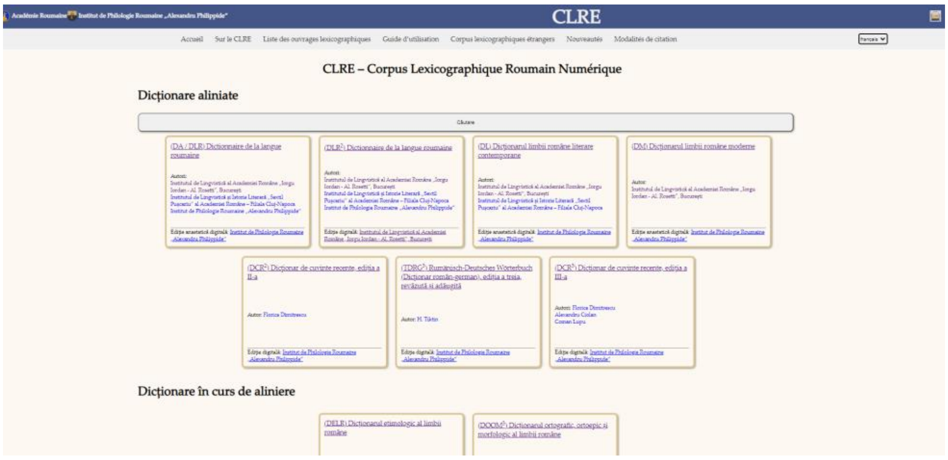
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1857 (Article)

2.2 CLRE. *Corpus lexicografic românesc electronic* [Corpus lexicographique roumain électronique]

CLRE est un projet fondamental de l'Académie Roumaine (commencé en 2014), réalisé par les chercheurs du Département de lexicologie et lexicographie de L'Institut de Philologie Roumaine «Alexandru Philippide», L'Académie Roumaine — Filiale de Iaşi (https://clre.solirom.ro/).

Le corpus représente une collection d'éditions numériques des dictionnaires les plus représentatifs de la langue roumaine de tous les temps, alignés au niveau d'entrée et au niveau de la catégorie lexico-grammaticale.

En mai 2022, la première édition numérique anastatique au niveau d'entrée et de la catégorie lexico-grammaticale du *Dictionnaire thésaurus de la langue roumaine*, édité sous les auspices de l'Académie roumaine (1913–2010), a été publiée dans son intégralité en CLRE.


Académie Roumaine Institut de Philologie Roumaine „Alexandru Philippide"
CLRE
Accueil
Sur le CLRE
Liste des ouvrages lexicographiques
Guide d'utilisation
Corpus lexicographiques étrangers
Nouveautés
Modalités de citation
français
Căutare

CLRE – Corpus Lexicographique Roumain Numérique

Dicționare aliniate

(DA / DLR) Dictionnaire de la langue roumaine

(DLR²) Dictionnaire de la langue roumaine

(DL) Dicționarul limbii române literare contemporane

(DM) Dicționarul limbii române moderne

(DCR²) Dicționar de cuvinte recente, ediția a II-a
Autor: Florica Dimitrescu

(TDRG³) Rumänisch-Deutsches Wörterbuch (Dicționar român-german), ediția a treia, revăzută și adăugită
Autor: H. Tiktin

(DCR³) Dicționar de cuvinte recente, ediția a III-a
Autori: Florica Dimitrescu
Alexandru Ciolan
Coman Lupu

Dicționare în curs de aliniere

(DELR) Dicționarul etimologic al limbii române

(DOOM³) Dicționarul ortografic, ortoepic și morfologic al limbii române

Figure 3: La première page de CLRE

Les principaux objectifs du projet CLRE sont:

- la création du plus grand corpus diachronique numérique de dictionnaires de langue roumaine, aligné, pour l'instant, au niveau de l'entrée, un corpus composé d'œuvres lexicographiques numérisées de la bibliographie DLR (transposés du format classique, sur papier, au format numérique) et à partir de dictionnaires numérisés (créés directement en format numérique éditable);
- la mise en libre accès pour le grand public des informations du CLRE. Partant du fait que la politique linguistique en Roumanie commence à prendre un contour plus ferme, les bons moyens ont été trouvés pour

pouvoir offrir un accès gratuit aux informations du corpus à tous ceux qui s'intéressent aux œuvres lexicographiques en langue roumaine;

— promotion des travaux lexicographiques réalisés sous les auspices de l'Académie Roumaine.

CLRE comprend différentes catégories de dictionnaires: généraux, explicatifs, étymologiques, encyclopédiques ou spécialisés, choisis selon le critère de leur importance pour la perspective diachronique de la langue roumaine.

À présent, le corpus CLRE, en développement continu, comprend 62 dictionnaires totalisant 124 volumes à divers stades de numérisation, dans laquelle chaque page ou image se voit attribuer des entrées annotées en lemmes ou variantes et des métadonnées sont attachées à chaque page ou image (liée à l'œuvre dont elle fait partie, numéro de page, type de page — pages précédentes, table des matières, pages arrière).

Chaque œuvre lexicographique dans CLRE est une édition anastatique, au sens de reproduction fidèle par numérisation des images des pages, la différence spécifique étant l'intégration de la collection d'images résultante dans une structure XML qui comprend également des métadonnées sur l'œuvre, des informations sur chaque mot-titre et des ajouts ou des corrections annotés (faits par les éditeurs CLRE)[11].

Par la conception, par la similarité en termes d'approche technique et par les dimensions, CLRE, le corpus lexicographique électronique roumain, peut être comparé à d'autres corpus lexicographiques européens:

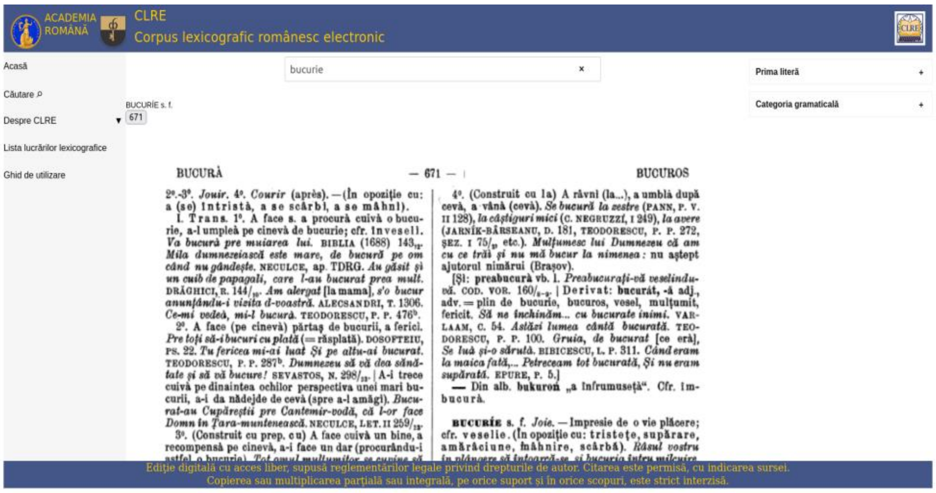
— *Das Wörterbuchnetz* — une collection de 37 dictionnaires au format numérique, créée à l'Université de Trier en Allemagne (https://www.woerterbuchnetz.de/);

— *Diccionarios de la lengua española* — la base de données contenant les dictionnaires édités et publiés par la Real Academia Española (https://www.rae.es/obras-academicas/diccionarios);

— le corpus lexicographique pour la langue française réalisé au cadre d'*Ortolang. Outils et ressources pour un traitement optimisé de la Langue*, par le Centre National de Ressources Textuelles et Lexicales — https://cnrtl.fr/dictionnaires/modernes/, etc.

L'interface CLRE permet actuellement plusieurs types de recherche (par entrée, selon la première lettre ou selon la catégorie grammaticale), selon différents critères, comme on peut le voir sur les figures ci-dessous:

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1857 (Article)



BUCURÀ — 671 — BUCUROS

2°.-3°. *Jouir.* 4°. *Courir* (après). — (În opoziție cu: a (se) întristà, a se scârbì, a se mâhnì).

I. Trans. 1°. A face s. a procurà cuivà o bucurie, a-l umpleà pe cinevà de bucurie; cfr. înveselì. *Va bucurà pre muiarea lui.* BIBLIA (1688) 143_{12}. *Mila dumnezeiască este mare, de bucură pe om când nu gândește.* NECULCE, ap. TDRG. *Au găsit și un cuib de papagali, care l-au bucurat prea mult.* DRĂGHICI, R. $144/_{10}$. *Am alergat* [la mama], *s'o bucur anunțându-i vizita d-voastră.* ALECSANDRI, T. 1306. *Ce-mi vedeà, mi-l bucurà.* TEODORESCU, P. P. 476^b.

2°. A face (pe cinevà) părtaș de bucurii, a fericì. *Pre toți să-i bucuri cu plată* (= răsplată). DOSOFTEIU, PS. 22. *Tu fericea mi-ai luat Și pe altu-ai bucurat.* TEODORESCU, P. P. 287^b. *Dumnezeu să vă dea sănătate și să vă bucure!* SEVASTOS, N. $298/_{33}$. | A-i trece cuivà pe dinaintea ochilor perspectiva unei mari bucurii, a-i da nădejde de cevà (spre a-l amăgì). *Bucurat-au Cupăreștii pre Cantemir-vodă, că l-or face Domn în Țara-muntenească.* NECULCE, LET. II $259/_{13}$.

3°. (Construit cu prep. cu) A face cuivà un bine, a recompensà pe cinevà, a-i face un dar (procurându-i

4°. (Construit cu la) A râvnì (la...), a umblà după cevà, a vânà (cevà). *Se bucură la zestre* (PANN, P. V. II 128), *la câștiguri mici* (C. NEGRUZZI, I 249), *la avere* (JARNÍK-BÂRSEANU, D. 181, TEODORESCU, P. P. 272, ȘEZ. I $75/_{9}$, etc.). *Mulțumesc lui Dumnezeu că am cu ce trăì și nu mă bucur la nimenea*: nu aștept ajutorul nimărui (Brașov).

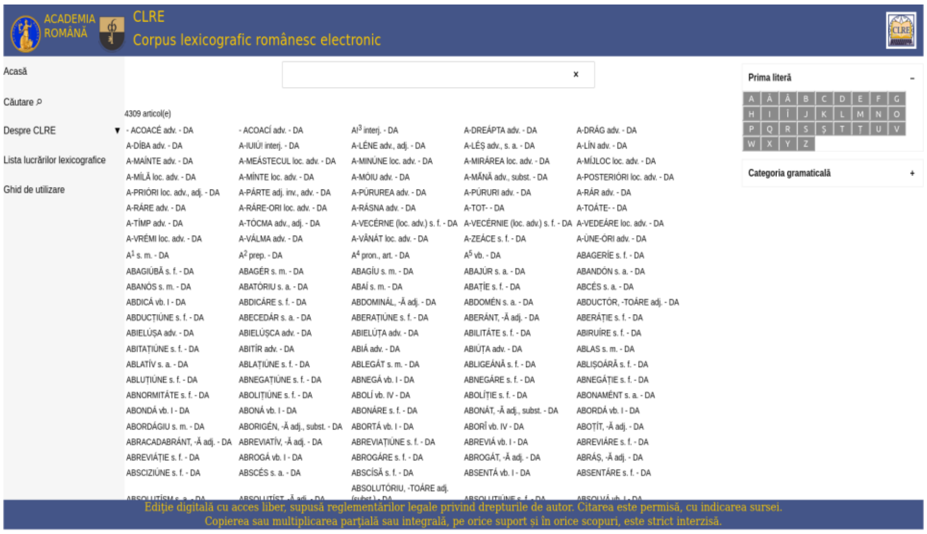
[Și: **preabucurà** vb. I. *Preabucurați-vă veselindu-vă.* COD. VOR. $160/_{8-9}$. | Derivat: **bucurát, -ă** adj., adv. = plin de bucurie, bucuros, vesel, mulțumit, fericit. *Să ne închinăm... cu bucurate inimi.* VARLAAM, C. 54. *Astăzi lumea cântă bucurată.* TEODORESCU, P. P. 100. *Gruia, de bucurat* [ce erà], *Se luà și-o sărutà.* BIBICESCU, L. P. 311. *Când eram la maica fată,... Petreceam tot bucurată, Și nu eram supărată.* EPURE, P. 5.]

— Din alb. **bukuron** „a înfrumusețà". Cfr. îmbucurà.

BUCURÍE s. f. *Joie.* — Impresie de o vie plăcere; cfr. veselie. (În opoziție cu: tristețe, supărare, amărăciune, mâhnire, scârbă). *Râsul vostru*

Ediție digitală cu acces liber, supusă reglementărilor legale privind drepturile de autor. Citarea este permisă, cu indicarea sursei.
Copierea sau multiplicarea parțială sau integrală, pe orice suport și în orice scopuri, este strict interzisă.

Figure 4: CLRE — le résultat de la recherche selon l'entrée

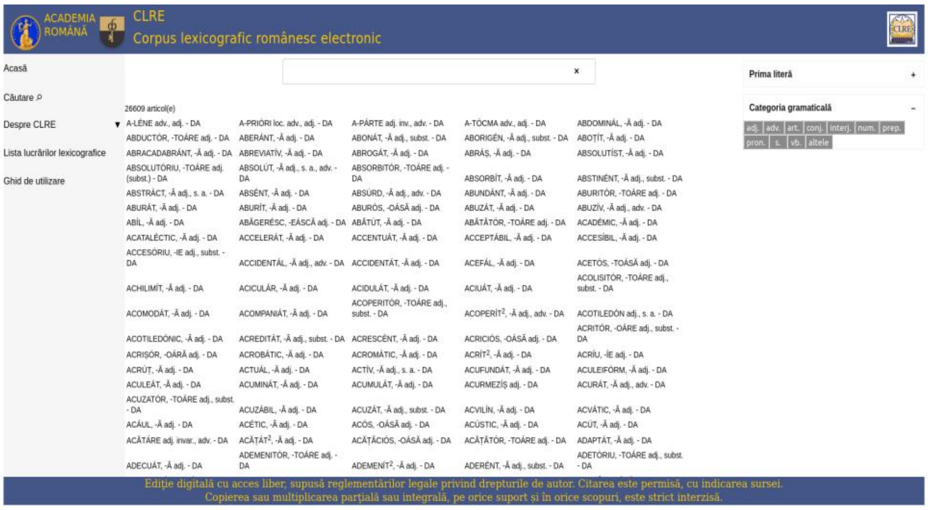


4309 articol(e)

- ACOACÉ adv. - DA	- ACOACÍ adv. - DA	A!³ interj. - DA	A-DREÁPTA adv. - DA	A-DRÁG adv. - DA
A-DÍBA adv. - DA	A-IUIÚ! interj. - DA	A-LÉNE adv., adj. - DA	A-LÉȘ adv., s. a. - DA	A-LÍN adv. - DA
A-MAÍNTE adv. - DA	A-MEÁSTECUL loc. adv. - DA	A-MINÚNE loc. adv. - DA	A-MIRÁREA loc. adv. - DA	A-MÍJLOC loc. adv. - DA
A-MÍLĂ loc. adv. - DA	A-MÍNTE loc. adv. - DA	A-MÓIU adv. - DA	A-MĂNĂ adv., subst. - DA	A-POSTERIÓRI loc. adv. - DA
A-PRIÓRI loc. adv., adj. - DA	A-PÁRTE adj. inv., adv. - DA	A-PÚRUREA adv. - DA	A-PÚRURI adv. - DA	A-RÁR adv. - DA
A-RÁRE adv. - DA	A-RÁRE-ORI loc. adv. - DA	A-RÁSNA adv. - DA	A-TOT- - DA	A-TOÁTE- - DA
A-TÍMP adv. - DA	A-TÓCMA adv., adj. - DA	A-VECÉRNE (loc. adv.) s. f. - DA	A-VECÉRNIE (loc. adv.) s. f. - DA	A-VEDEÁRE loc. adv. - DA
A-VRÉMI loc. adv. - DA	A-VÁLMA adv. - DA	A-VÂNÁT loc. adv. - DA	A-ZEÁCE s. f. - DA	A-ÚNE-ÓRI adv. - DA
A¹ s. m. - DA	A² prep. - DA	A⁴ pron., art. - DA	A⁵ vb. - DA	ABAGERÍE s. f. - DA
ABAGIÚBĂ s. f. - DA	ABAGÉR s. m. - DA	ABAGÍU s. m. - DA	ABAJÚR s. a. - DA	ABANDÓN s. a. - DA
ABANÓS s. m. - DA	ABATÓRIU s. a. - DA	ABAÍ s. m. - DA	ABAȚÍE s. f. - DA	ABCÉS s. a. - DA
ABDICÁ vb. I - DA	ABDICÁRE s. f. - DA	ABDOMINÁL, -Ă adj. - DA	ABDOMÉN s. a. - DA	ABDUCTÓR, -TOÁRE adj. - DA
ABDUCȚIÚNE s. f. - DA	ABECEDÁR s. a. - DA	ABERAȚIÚNE s. f. - DA	ABERÁNT, -Ă adj. - DA	ABERÁȚIE s. f. - DA
ABIELÚȘA adv. - DA	ABIELÚȘCA adv. - DA	ABIELÚȚA adv. - DA	ABILITÁTE s. f. - DA	ABIRUÍRE s. f. - DA
ABITAȚIÚNE s. f. - DA	ABITÍR adv. - DA	ABIÁ adv. - DA	ABIÚȚA adv. - DA	ABLAS s. m. - DA
ABLATÍV s. a. - DA	ABLAȚIÚNE s. f. - DA	ABLEGÁT s. m. - DA	ABLIGEÁNĂ s. f. - DA	ABLIȘOÁRĂ s. f. - DA
ABLUȚIÚNE s. f. - DA	ABNEGAȚIÚNE s. f. - DA	ABNEGÁ vb. I - DA	ABNEGÁRE s. f. - DA	ABNEGÁȚIE s. f. - DA
ABNORMITÁTE s. f. - DA	ABOLIȚIÚNE s. f. - DA	ABOLÍ vb. IV - DA	ABOLÍȚIE s. f. - DA	ABONAMÉNT s. a. - DA
ABONDÁ vb. I - DA	ABONÁ vb. I - DA	ABONÁRE s. f. - DA	ABONÁT, -Ă adj., subst. - DA	ABORDÁ vb. I - DA
ABORDÁGIU s. m. - DA	ABORIGÉN, -Ă adj., subst. - DA	ABORTÁ vb. I - DA	ABORÍ vb. IV - DA	ABOȚÍT, -Ă adj. - DA
ABRACADABRÁNT, -Ă adj. - DA	ABREVIATÍV, -Ă adj. - DA	ABREVIAȚIÚNE s. f. - DA	ABREVIÁ vb. I - DA	ABREVIÁRE s. f. - DA
ABREVIÁȚIE s. f. - DA	ABROGÁ vb. I - DA	ABROGÁRE s. f. - DA	ABROGÁT, -Ă adj. - DA	ABRÁȘ, -Ă adj. - DA
ABSCIZIÚNE s. f. - DA	ABSCÉS s. a. - DA	ABSCÍSĂ s. f. - DA	ABSENTÁ vb. I - DA	ABSENTÁRE s. f. - DA

Ediție digitală cu acces liber, supusă reglementărilor legale privind drepturile de autor. Citarea este permisă, cu indicarea sursei.
Copierea sau multiplicarea parțială sau integrală, pe orice suport și în orice scopuri, este strict interzisă.

Figure 5: CLRE — le résultat de la recherche selon la première lettre



26609 articol(e)

A-LÉNE adv., adj. - DA	A-PRIÓRI loc. adv., adj. - DA	A-PÁRTE adj. inv., adv. - DA	A-TÓCMA adv., adj. - DA	ABDOMINÁL, -Ă adj. - DA
ABDUCTÓR, -TOÁRE adj. - DA	ABERÁNT, -Ă adj. - DA	ABONÁT, -Ă adj., subst. - DA	ABORIGÉN, -Ă adj., subst. - DA	ABOȚÍT, -Ă adj. - DA
ABRACADABRÁNT, -Ă adj. - DA	ABREVIATÍV, -Ă adj. - DA	ABROGÁT, -Ă adj. - DA	ABRÁȘ, -Ă adj. - DA	ABSOLUTÍST, -Ă adj. - DA
ABSOLUTÓRIU, -TOÁRE adj. (subst.) - DA	ABSOLÚT, -Ă adj., s. a., adv. - DA	ABSORBITÓR, -TOÁRE adj. - DA	ABSORBÍT, -Ă adj. - DA	ABSTINÉNT, -Ă adj., subst. - DA
ABSTRÁCT, -Ă adj., s. a. - DA	ABSÉNT, -Ă adj. - DA	ABSÚRD, -Ă adj., adv. - DA	ABUNDÁNT, -Ă adj. - DA	ABURITÓR, -TOÁRE adj. - DA
ABURÁT, -Ă adj. - DA	ABURÍT, -Ă adj. - DA	ABURÓS, -OÁSĂ adj. - DA	ABUZÁT, -Ă adj. - DA	ABUZÍV, -Ă adj., adv. - DA
ABÍL, -Ă adj. - DA	ABĂGERÉSC, -EÁSCĂ adj. - DA	ABĂTÚT, -Ă adj. - DA	ABĂTĂTÓR, -TOÁRE adj. - DA	ACADÉMIC, -Ă adj. - DA
ACATALÉCTIC, -Ă adj. - DA	ACCELERÁT, -Ă adj. - DA	ACCENTUÁT, -Ă adj. - DA	ACCEPTÁBIL, -Ă adj. - DA	ACCESÍBIL, -Ă adj. - DA
ACCESÓRIU, -IE adj., subst. - DA	ACCIDENTÁL, -Ă adj., adv. - DA	ACCIDENTÁT, -Ă adj. - DA	ACEFÁL, -Ă adj. - DA	ACETÓS, -TOÁSĂ adj. - DA
ACHILIMÍT, -Ă adj. - DA	ACICULÁR, -Ă adj. - DA	ACIDULÁT, -Ă adj. - DA	ACIUÁT, -Ă adj. - DA	ACOLISITÓR, -TOÁRE adj., subst. - DA
ACOMODÁT, -Ă adj. - DA	ACOMPANIÁT, -Ă adj. - DA	ACOPERITÓR, -TOÁRE adj., subst. - DA	ACOPERÍT^2, -Ă adj., adv. - DA	ACOTILEDÓN adj., s. a. - DA
ACOTILEDÓNIC, -Ă adj. - DA	ACREDITÁT, -Ă adj., subst. - DA	ACRESCÉNT, -Ă adj. - DA	ACRICIÓS, -OÁSĂ adj. - DA	ACRITÓR, -OÁRE adj., subst. - DA
ACRIȘÓR, -OÁRĂ adj. - DA	ACROBÁTIC, -Ă adj. - DA	ACROMÁTIC, -Ă adj. - DA	ACRÍT^2, -Ă adj. - DA	ACRÍU, -ÍE adj. - DA
ACRÚȚ, -Ă adj. - DA	ACTUÁL, -Ă adj. - DA	ACTÍV, -Ă adj., s. a. - DA	ACUFUNDÁT, -Ă adj. - DA	ACULEIFÓRM, -Ă adj. - DA
ACULEÁT, -Ă adj. - DA	ACUMINÁT, -Ă adj. - DA	ACUMULÁT, -Ă adj. - DA	ACURMEZÍȘ adj. - DA	ACURÁT, -Ă adj., adv. - DA
ACUZATÓR, -TOÁRE adj., subst. - DA	ACUZÁBIL, -Ă adj. - DA	ACUZÁT, -Ă adj., subst. - DA	ACVILÍN, -Ă adj. - DA	ACVÁTIC, -Ă adj. - DA
ACÁUL, -Ă adj. - DA	ACÉTIC, -Ă adj. - DA	ACÓS, -OÁSĂ adj. - DA	ACÚSTIC, -Ă adj. - DA	ACÚT, -Ă adj. - DA
ACĂTÁRE adj. invar., adv. - DA	ACĂȚÁT^2, -Ă adj. - DA	ACĂȚĂCIÓS, -OÁSĂ adj. - DA	ACĂȚĂTÓR, -TOÁRE adj. - DA	ADAPTÁT, -Ă adj. - DA
ADECUÁT, -Ă adj. - DA	ADEMENITÓR, -TOÁRE adj. - DA	ADEMENÍT^2, -Ă adj. - DA	ADERÉNT, -Ă adj., subst. - DA	ADETÓRIU, -TOÁRE adj., subst. - DA

Ediție digitală cu acces liber, supusă reglementărilor legale privind drepturile de autor. Citarea este permisă, cu indicarea sursei.
Copierea sau multiplicarea parțială sau integrală, pe orice suport și în orice scopuri, este strict interzisă.

Figure 6: CLRE — le résultat de la recherche selon la catégorie grammaticale

Au fur et à mesure que le corpus se développera et que plusieurs ouvrages lexicographiques en éditions numériques, anastatiques ou créées directement en version digitale, seront accessibles au public, de nouvelles facilités de consultation seront créées pour le public-cible. Une étape essentielle à mettre en œuvre est l'alignement avec eDTLR (lorsque les résultats finaux seront accessibles) et, bien sûr, DLR^2.

L'utilisation du CLRE pour la rédaction du DLR^2 et pour d'autres projets lexicographiques et la corrélation avec d'autres ressources linguistiques ou multimédias, amènent la lexicographie roumaine à un niveau comparable à la lexicographie européenne (par exemple, avec *DÉRom. Dictionnaire Étymologique Roman* — www.atilf.fr/DERom/) ou permet l'inclusion de la lexicographie roumaine dans des réseaux internationales de lexicographie, telle, par exemple, *EneL. European Network of e-Lexicography* (www.elexicography.eu), *ELEXIS. European Lexicographic Infrastructure* (https://elex.is/).

2.3 D'autres projets lexicographiques numériques dans les instituts de l'Académie Roumaine

DELR. Dicționarul etimologic al limbii române [Le dictionnaire étymologique de la langue roumaine] est l'ouvrage lexicographique le plus complet de ce type pour la langue roumaine, comprenant à la fois les mots utilisés dans la langue standard actuelle et ceux de l'ancien roumain, qui ont disparu ou sont devenus archaïques, à côté des régionalismes et de ceux utilisés dans les langages spécialisés.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1857 (Article)

DELR est en cours de développement à l'Institut de Linguistique de l'Académie roumaine «Iorgu Iordan — Al. Rosetti» de Bucarest.

Actuellement DELR comprend les lettres A, C et la première partie de la lettre D (jusqu'au *djinn*), totalisant 30 514 mots (avec 15 251 variantes), regroupés en 8 809 nids lexicaux. DELR est publié en ligne à l'adresse: https://delr.lingv.ro/.

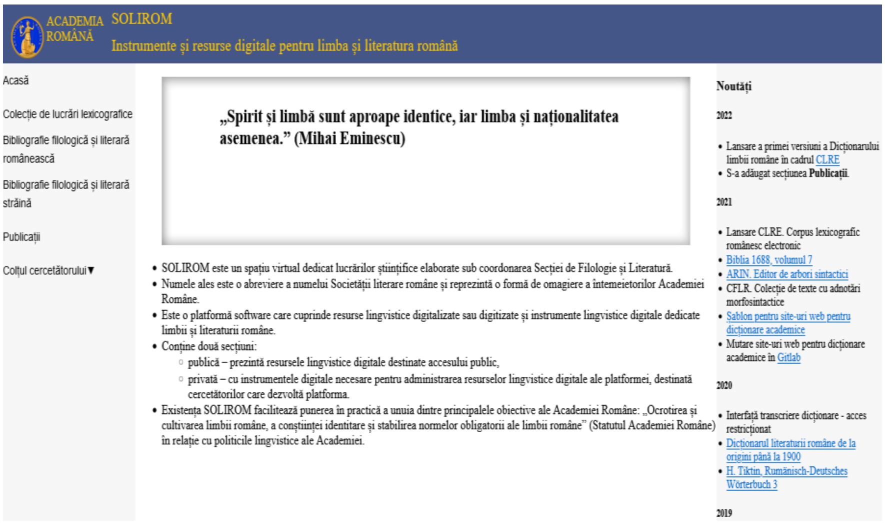
2.4 SOLIROM — instrument pour les ressources et les outils en format numérique pour la langue et la littérature roumaine

SOLIROM représente une plate-forme logicielle de l'Académie Roumaine, qui comprend des ressources linguistiques numériques et des outils linguistiques dédiés à la langue et à la littérature roumaines (DLR[2], CLRE, eDTLR, TDRG[12], etc.) créés dans les instituts de philologie subordonnés à la Section de philologie et littérature de l'Académie roumaine (https://solirom.ro/).

SOLIROM publie des résultats concernant la langue et la littérature roumaines; en outre, la plate-forme fournit aux chercheurs des outils et des ressources pour la recherche numérique, les informations étant visibles en deux sections:

— publique — on présente des ressources linguistiques numériques accessibles au public;

— privée — avec les outils numériques nécessaires à la gestion des ressources linguistiques numériques de la plate-forme, pour les chercheurs qui la développement.


ACADEMIA ROMÂNĂ
SOLIROM
Instrumente și resurse digitale pentru limba și literatura română
Acasă
Colecție de lucrări lexicografice
Bibliografie filologică și literară românească
Bibliografie filologică și literară străină
Publicații
Colțul cercetătorului▼
Noutăți

„Spirit și limbă sunt aproape identice, iar limba și naționalitatea asemenea." (Mihai Eminescu)

- SOLIROM este un spațiu virtual dedicat lucrărilor științifice elaborate sub coordonarea Secției de Filologie și Literatură.
- Numele ales este o abreviere a numelui Societății literare române și reprezintă o formă de omagiere a întemeietorilor Academiei Române.
- Este o platformă software care cuprinde resurse lingvistice digitalizate sau digitizate și instrumente lingvistice digitale dedicate limbii și literaturii române.
- Conține două secțiuni:
  - publică – prezintă resursele lingvistice digitale destinate accesului public,
  - privată – cu instrumentele digitale necesare pentru administrarea resurselor lingvistice digitale ale platformei, destinată cercetătorilor care dezvoltă platforma.
- Existența SOLIROM facilitează punerea în practică a unuia dintre principalele obiective ale Academiei Române: „Ocrotirea și cultivarea limbii române, a conștiinței identitare și stabilirea normelor obligatorii ale limbii române" (Statutul Academiei Române) în relație cu politicile lingvistice ale Academiei.

2022

- Lansare a primei versiuni a Dicționarului limbii române în cadrul CLRE
- S-a adăugat secțiunea **Publicații**.

2021

- Lansare CLRE. Corpus lexicografic românesc electronic
- Biblia 1688, volumul 7
- ARIN. Editor de arbori sintactici
- CFLR. Colecție de texte cu adnotări morfosintactice
- Șablon pentru site-uri web pentru dicționare academice
- Mutare site-uri web pentru dicționare academice în Gitlab

2020

- Interfață transcriere dicționare - acces restricționat
- Dicționarul literaturii române de la origini până la 1900
- H. Tiktin, Rumänisch-Deutsches Wörterbuch 3

2019

Figure 7: La première page du SOLIROM (https://solirom.ro/)

A présent l'Académie Roumaine a englobé dans sa nouvelle page électronique une zone spéciale dédiée aux ressources (numériques) en langue roumaine (des dictionnaires, la grammaire, etc.) — https://acad.ro/institutia/limba_romana.html.

2.5 CoRoLa. Corpus numérique pour la langue roumaine contemporaine

Les projets lexicographiques nécessitent aussi des corpus linguistiques à partir desquels les dictionnaires peuvent être rédigés.

Pour la langue roumaine contemporaine *CoRoLa. Corpus computațional de referință pentru limba română contemporană* [Corpus numérique de référence pour la langue roumaine contemporaine] représente un corpus linguistique, élaboré à partir de 2014 en tant que programme prioritaire de l'Académie Roumaine et qui contient divers textes, datant de 1989 à nos jours, le but de sa création étant de fournir une image objective de la langue roumaine écrite et parlée actuelle. Le corpus est ouvert au public à travers une interface de recherche de données textuelles et une de recherche de données audio.

Le corpus est consultable en ligne à l'adresse: http://corola.racai.ro/.

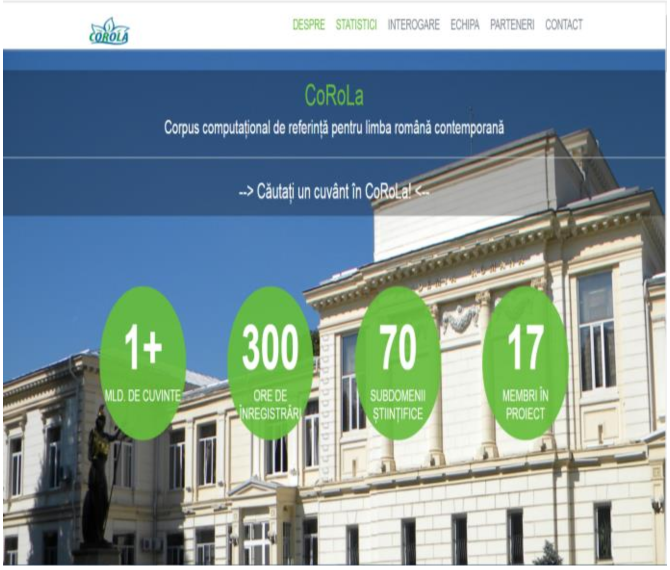

DESPRE STATISTICI INTEROGARE ECHIPA PARTENERI CONTACT
CoRoLa
Corpus computațional de referință pentru limba română contemporană
--> Căutați un cuvânt în CoRoLa! <--
1+
MLD. DE CUVINTE
300
ORE DE ÎNREGISTRĂRI
70
SUBDOMENII ȘTIINȚIFICE
17
MEMBRI ÎN PROIECT

Figure 8: La première page du CoRoLa

2.6 Autres ressources lexicographiques roumaines disponibles en ligne

2.6.1 eRomLex

Primele dicționare bilingve românești (secolul al XVII-lea). Corpus digital prelucrat și aliniat

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1857 (Article)

(eRomLex) [Les premiers dictionnaires bilingues roumains (XVIIe siècle). Corpus numérique traité et aligné (eRomLex)] — http://www.scriptadacoromanica.ro/bin/view/eRomLex/[13] — est un projet de (retro-)numérisation, bien inédit pour le roumain par la valorisation à travers une édition comparative numérisée des premiers efforts lexicographiques roumains des lexiques bilingues roumains, partant du fait qu'ils ont la même source, afin de mettre en évidence les caractéristiques de ce réseau lexicographique et de faciliter l'accès à leur contenu. Il s'agit de six lexiques bilingues slavons-roumains du XVIIe siècle (tous manuscrits), qui représentent des éléments très importants dans la première période de développement de la lexicographie roumaine. Ces 6 lexiques sont comparés à leur source commune (le lexique slavon-ruthène publié par Pamvo Berynda à Kiev, en 1627). Ce réseau lexicographique va être complété par des études concernant leur rapport à la source, les innovations à son égard, leurs usages probables, les filiations entre lexiques, le contexte socio-culturel, le but de cette compilation, leurs utilisations potentielles, etc.

Le projet propose également une méthode pour la mise en place d'une base de données alignée qui pourrait être connectée aux dictionnaires électroniques / corpus lexicographiques électroniques roumains déjà existants (CLRE, par exemple), complétant les informations linguistiques qu'ils contiennent et lancera une approche plus étendue destinée à produire des dictionnaires traités et annotés des corpus numérisés de roumain ancien auxquels ces lexiques pourraient être reliés afin de permettre la mise en valeur des pratiques de traduction, et les usages spécifiques du matériel linguistique; ces aspects pourraient éclairer certaines questions liées à l'origine formulées jusqu'à présent et qui sont difficiles à vérifier au moyen des méthodes classiques d'interrogation des textes.

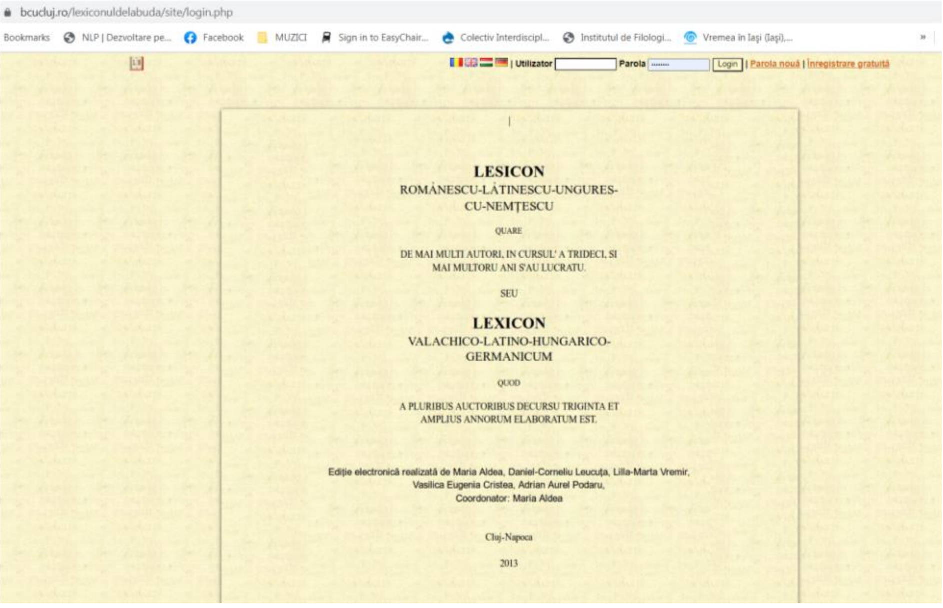
Les étapes de la réalisation de ce projet de numérisation supposent: la translittération en alphabet latin des définitions roumaines et la transcription des mots slavons; le traitement automatique du matériel (unification des graphies; élimination des diacritiques); l'alignement des entrées correspondantes; l'indexation et recherche multicritères; des études quantitatives: nombre d'entrées communes, entrées spécifiques à un certain lexique; des innovations par rapport à la source; la comparaison des inventaires lexicaux; des études quantitatives: utilité du matériel pour DLR — premières attestations, sens / formes inédit(e)s; (directions de recherches futures) alignement a d'autres ressources numériques (e.g. CLRE)[14].

Le corpus réalisé par ce projet complètera l'image d'ensemble des débuts de la lexicographie roumaine.

2.6.2 Lexiconul de la Buda [Lexique de Buda]

Une autre ressource lexicographique roumaine disponible en ligne est *Lexiconul de la Buda*[15], l'édition électronique du premier dictionnaire étymologique et explicatif de la langue roumaine et c'est une référence pour la lexicographie roumaine moderne. Pour le public qui consulte le dictionnaire en ligne, l'accès est gratuit après inscription sur le site, en créant un compte et un mot de passe. L'interface est extrêmement accessible, conçue pour être fonctionnelle, utile, en fonction des différents critères de requête, permettant ainsi un survol rapide du texte (en quatre langues: roumain, latin, hongrois et allemand). Ceci est le résultat d'un projet coordonné par l'Université Babeş-Bolyai, Cluj-Napoca.

Lexiconul de la Buda est publié en ligne à l'adresse http://bcucluj.ro/lexiconuldelabuda/site/login.php.


bcucluj.ro/lexiconuldelabuda/site/login.php
Bookmarks
NLP | Dezvoltare pe...
Facebook
MUZICI
Sign in to EasyChair...
Colectiv Interdiscipl...
Institutul de Filologi...
Vremea în Iaşi (Iaşi),...
| Utilizator
Parola
Login
| Parola nouă | Înregistrare gratuită
I

LESICON
ROMÂNESCU-LÂTINESCU-UNGURES-
CU-NEMŢESCU

QUARE

DE MAI MULTI AUTORI, IN CURSUL' A TRIDECI, SI
MAI MULTORU ANI S'AU LUCRATU.

SEU

LEXICON
VALACHICO-LATINO-HUNGARICO-
GERMANICUM

QUOD

A PLURIBUS AUCTORIBUS DECURSU TRIGINTA ET
AMPLIUS ANNORUM ELABORATUM EST.

Ediţie electronică realizată de Maria Aldea, Daniel-Corneliu Leucuţa, Lilla-Marta Vremir,
Vasilica Eugenia Cristea, Adrian Aurel Podaru,
Coordonator: Maria Aldea

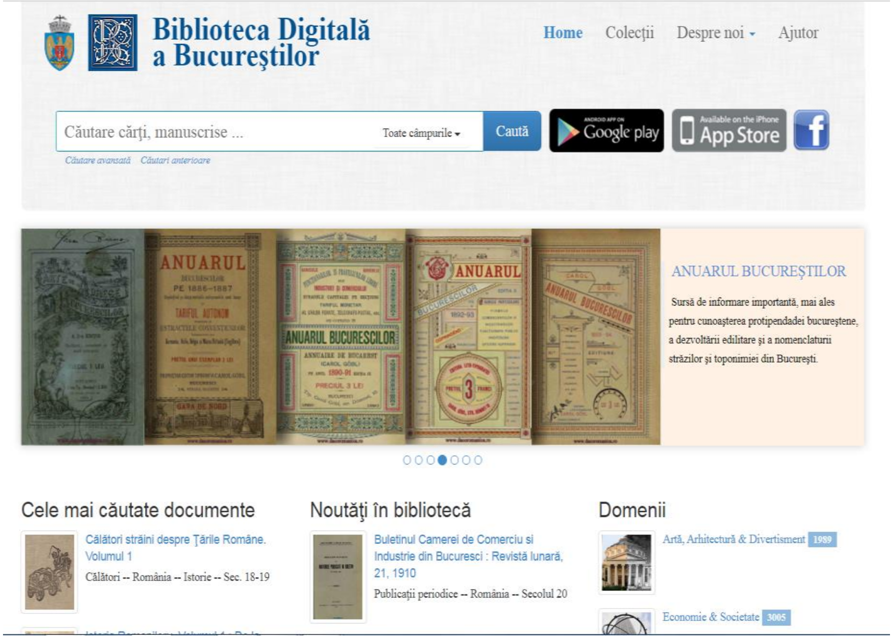
Cluj-Napoca

2013

Figure 9: La première page du *Lexiconul de la Buda*

2.6.3 DIGIBUC

DIGIBUC (http://www.digibuc.ro/) représente la plus grande bibliothèque numérique roumaine, étant un projet réalisé par la Bibliothèque métropolitaine de Bucarest et la Bibliothèque de l'Académie roumaine, et est le partenaire officiel d'EUROPEANA, la bibliothèque numérique la plus importante d'Europe — (http://www.europeana.eu/portal/).



ANUARUL BUCUREŞTILOR

Sursă de informare importantă, mai ales pentru cunoaşterea protipendadei bucureştene, a dezvoltării edilitare şi a nomenclaturii străzilor şi toponimiei din Bucureşti.

Călători străini despre Ţările Române. Volumul 1
Călători -- România -- Istorie -- Sec. 18-19

Buletinul Camerei de Comerciu si Industrie din Bucuresci : Revistă lunară, 21, 1910
Publicaţii periodice -- România -- Secolul 20

Figure 10: La première page du DIGIBUC

3. Perspectives actuelles dans la lexicographie roumaine

Les avantages de la numérisation de la recherche lexicographique roumaine résident dans le fait que les ressources créées représentent:

— des outils de travail pour les lexicographes et pour les linguistes en général;
— une source d'information facilement accessible pour ceux qui s'intéressent à la langue roumaine;
— base de création d'applications pour divers appareils nécessitant un lexique;
— des points de départ pour de futures recherches.

Ces résultats permettent également la corrélation avec d'autres ressources linguistiques ou multimédias. Ainsi, dans une prochaine étape, des connexions pourront être établies avec des informations issues de plateformes encyclopédiques telles que Wikipédia, une approche qui pourrait augmenter la visibilité des données lexicographiques incluses dans CLRE, par exemple, mais aussi enrichir les informations que le lecteur/l'utilisateur a besoin.

Donc, les perspectives de la lexicographie roumaine envisagent la création des dictionnaires numériques à partir d'un corpus textuel; le développement du corpus lexicographique CLRE; aligner les variantes numériques du *Dictionnaire-trésor de la langue roumaine* dans CLRE; rédiger / publier des dictionnaires

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1857 (Article)

en format hybride (en format classique / en papier et, en même temps, en format numérique) ou exclusivement en format classique.

4. Conclusions

L'objectif de cet article est de mettre en évidence, en général, l'état actuel de la recherche lexicographique en Roumanie et ses perspectives.

Les dictionnaires électroniques et les corpus de textes, structurés en bases de données, facilitent la connaissance, la préservation et le maintien de l'identité culturelle au niveau linguistique et permettent l'inclusion d'une langue nationale dans le domaine d'intérêt de la recherche numérique des langues naturelles, au niveau mondial.

La numérisation de la recherche lexicographique roumaine est en plein développement, l'étape actuelle de l'évolution de la lexicographie roumaine impliquant, d'une part, la numérisation des ressources existantes et, d'autre part, la création de dictionnaires, de nouvelles ressources et instruments, directement en format électronique.

A travers tous les projets de numérisation de la recherche lexicographique roumaine, la langue roumaine entre plus solidement dans le circuit d'une communication plus facile.

Le processus de numérisation place la lexicographie académique roumaine à un niveau comparable à la lexicographie internationale et permet la connexion avec des projets lexicographiques de l'étranger et l'inclusion de la Roumanie dans la sphère d'intérêt des grands réseaux lexicographiques internationaux

Remerciements

Le présent article envisage aussi les résultats du projet de recherche *Primele dicționare bilingve românești (secolul al XVII-lea). Corpus digital prelucrat și aliniat* [Les premiers dictionnaires roumains bilingues (XVIIe siècle). Corpus numériques traités et alignés] http://www.scriptadacoromanica.ro/bin/view/eRomLex/), subventionné par le Ministère de la Recherche, de l'Innovation et de la Numérisation de Roumanie, CNCS/CCCDI — UEFISCDI, projet PNIII P11.1TE 20190517.

Endnotes

1. Pour une présentation détaillée du rôle des instituts de l'Académie Roumaine dans le processus de numérisation, voir Tamba 2022b.
2. Ainsi, dans cet article on va détailler aussi les étapes d'un projet de recherche intitulé *Primele dicționare bilingve românești (secolul al XVII-lea). Corpus digital prelucrat și aliniat (eRomLex)* [Les premiers dictionnaires bilingues roumains (XVIIe siècle). Corpus numérique traité et aligné (eRomLex)] réalisé à l'Université «Alexandru Ioan Cuza», de Iași (http://www.scriptadacoromanica.ro/bin/view/eRomLex/).

3. Pour exemple, https://dexonline.ro/ — une plate-forme comprenant bien des dictionnaires de langue roumaine, projet initiée par des volontaires.
4. *Dicționarul limbii române* — appelé aussi *Dicționarul tezaur al limbii române* [Le Dictionnaire-trésor de la langue roumaine] ou *Dicționarul Academiei* [Le Dictionnaire de l'Académie] — est rédigé dans les Départements de Lexicographie des instituts de langue et de littérature roumaine de l'Académie Roumaine: Institutul de Lingvistică "Iorgu Iordan — Alexandru Rosetti" (ILB), București — https://lingv.ro/; Institutul de Filologie Română "A. Philippide" (IFR), Iași — https://philippide.ro/; Institutul de Lingvistică și Istorie Literară "Sextil Pușcariu" (ILIL), Cluj-Napoca — http://inst-puscariu.ro/.
5. Toutes les informations concernant le projet CLRE seront présentées plus en détail dans le sous-chapitre suivant.
6. Image extrait du Tamba (2017b: 141).
7. Pour plus de détails sur les projets réalisés au cours de la période 2003–2017, voir Haja (2017), Tamba (2017b), Haja et Tamba (2022).
8. Pour plus de détails sur ce projet, voir Haja et al. (2005).
9. Au cours des deux dernières années, au cadre de l'Académie roumaine — Filiale de Iași se déroulent un autre projet (financé au niveau national par compétition), intéressant du point de vue de la valorisation des informations de la première édition du DLR, mais aussi du point de vue de la création de ressources numérisée. Le projet *TAFOC — Terminologia astronomică românească: științific vs popular. Fenomene, obiecte cosmice și constelații* [La terminologie astronomique roumaine: scientifique vs. populaire). Des phénomènes, des objets cosmiques et des constellations] (directrice du projet: CS I dr. Cristina Michaela Florescu; https://tafoc.solirom.ro/) propose une systématisation linguistique de la terminologie scientifique et populaire / commune de l'astronomie (champ lexical qui n'a pas été étudié systématiquement jusqu'à présent, tant en roumain, que dans les autres langues romanes), l'analyse ayant un caractère diachronique prononcé, en suivant l'étymologie et l'évolution historique des termes.
10. Toutes les informations concernant la plate-forme SOLIROM seront présentées plus en détail dans un autre sous-chapitre de cet article.
11. Pour une présentation détaillée sur CLRE, voir Haja et Tamba (2022), Tamba (2022a).
12. La version numérique de TDRG — H. Tiktin, *Rumänisch-Deutsches Wörterbuch* (la première édition: 1896–1926). La troisième édition de ce dictionnaire (publiée de 2003 à 2005) a été numérisée, sous le modèle de l'eDTLR, lors de la collaboration entre l'Albert-Ludwigs-Universität de Fribourg, en Allemagne, et l'Académie Roumaine. TDRG est publié en ligne à l'adresse: https://tdrg.solirom.ro/.
13. Pour plus de détails sur ce projet, voir Gînsac et al. (2022).
14. Pour une présentation détaillée sur eRomLex, voir Haja et Tamba (2022), Tamba (2022a).
15. *Lesicon romanescu–latinescu–ungurescu–nemtescu quare de mai mulţi autori, in cursul a trideci, si mai multoru ani s'au lucrat. Seu Lexicon valachico–latino–hungarico–germanicum quod a pluribus auctoribus decursu triginta et amplius annorum elaboratum est*. Budae, Typis et Sumtibus Typografiae Regiae Universitatis Hungaricae, 1825. Ediție electronică de Maria Aldea, Daniel-Corneliu Leucuţa, Lilla-Marta Vremir, Vasilica Eugenia Cristea și Adrian Aurel Podaru, Cluj-Napoca, 2013; http://bcucluj.ro/lexiconuldelabuda/site/login.php.

Références

Littérature secondaire

Abdelzaher, Esra M. 2022. An Investigation of Corpus Contributions to Lexicographic Challenges over the Past Ten Years. *Lexikos* 32: 162-179.

Busuioc, M.M., N. Mihai et Al. Anghelina. 2018. Lexicografie și lexicologie. Sala, M. et N. Saramandu. 2018. *Lingvistica românească:* 247-290. București: Editura Academiei Române.

Clim, M.-R., E. Tamba, A. Catană-Spenchiu et M. Patraşcu. 2016. *CLRE. Corpus lexicographique roumain essentiel.* 100 dictionnaires de la langue roumaine alignés au niveau de l'entrée et, partiellement, au niveau du sens. Éva Buchi, Jean-Paul Chauveau, Jean-Marie Pierrel (Éds.). 2016. *Actes du XXVIIe Congrès international de linguistique et de philologie romanes (Nancy, 15-20 juillet 2013). Vol. 2, Section 16:* 1611-1622. Strasbourg: ÉLiPhi.
URL: http://www.atilf.fr/cilpr2013/actes/section-16.html

Ernst, G. 2013. Romanian. Gouws, R.H., U. Heid, W. Schweickard et H.E. Wiegand (Éds.). 2013. *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Special Focus on Computational Lexicography:* 687-701. Berlin/Boston: De Gruyter.

Gînsac, A.-M., M.-A. Moruz et M. Ungureanu. 2022. The First Romanian Dictionaries (17th century). Digital Aligned Corpus. Klosa-Kückelhaus, A., S. Engelberg, C. Möhrs et P. Storjohann (Éds.). 2022. *Dictionaries and Society. Proceedings of the XX EURALEX International Congress, 12–16 July 2022, Mannheim, Germany*: 222-229. Mannheim: IDS-Verlag.

Gînsac, A.-M. et M. Ungureanu. 2018. La lexicographie slavonne-roumaine au XVIIe siècle. Adaptations roumaines d'après *le Leksikon slavenorosskij* de Pamvo Berynda. *Zeitschrift für romanische Philologie* 134(3): 845-876.

Haja, G. (Éd.). 2017. *Lexicografia academică românească. Studii. Proiecte.* Iași: Editura Universității "Alexandru Ioan Cuza".

Haja, G., E. Dănilă, C. Forăscu et B.-M. Aldea. 2005. *Dicționarul limbii române (DLR) în format electronic. Studii privind achiziționarea.* Iași: Editura Alfa.

Haja, G. et E.I. Tamba. 2022. Publicarea online a primului volum din *CLRE. Corpus lexicografic românesc electronic* — contextualizare. Clim, M.-R. et E.I. Tamba (Coord.). 2022. *Actele Colocviului internațional "Lexicografia academică românească. Provocările informatizării", 2020–2021:* 213-218. Cluj-Napoca: Presa Universitară Clujeană.
URL: http://www.editura.ubbcluj.ro/bd/ebooks/pdf/3188.pdf

Haja, G., E.I. Tamba, M.-R. Clim, C. Teodorescu et A. Anghelina. 2020. Metode și tehnici actuale în redactarea *Dicționarului limbii române informatizat (DLRi).* Ichim, Ofelia (Éd.). 2020. *România în spaţiul euroatlantic: interferenţe culturale şi lingvistice*: 139-160. București: Editura Tracus Arte.

Hanks, P. 2013. Lexicography from Earliest Times to the Present. Allan, K. 2013. *The Oxford Handbook of the History of Linguistics*: 503-536. Oxford: Oxford University Press.

Hartmann, R.R.K. et G. James. 1998. *Dictionary of Lexicography.* London: Routledge.

Kirchmeier, S. 2020. Trends in European Language Policies with a View to Language Technology. *Standard Language/ Bendrinė Kalba* 93(2020): 1-23.
URL: http://journals.lki.lt/bendrinekalba

Moruz, M.-A. et M. Ungureanu. 2022. 17th-Century Romanian Lexical Resources and their Influence on Romanian Written Tradition. Klosa-Kückelhaus, A., S. Engelberg, C. Möhrs et P. Storjohann (Éds.). 2022. *Dictionaries and Society. Proceedings of the XX EURALEX International Congress, 12–16 July 2022, Mannheim, Germany*: 745-754. Mannheim: IDS-Verlag.

Tamba, E. 2014. La lexicografía Rumana. Historia y Actualidad. Córdoba Rodríguez, F., E. González Seoane et María Dolores Sánchez Palomino *Lexicografía de las lenguas románicas:* 265-282. *Perspectiva histórica. Vol. I.* Berlin/Boston: De Gruyter.

Tamba, Elena Isabelle. 2017a. CLRE. *Corpus lexicografic românesc esenţial.* 100 de dicţionare din Bibliografia DLR aliniate la nivel de intrare şi la nivel de sens. Haja, Gabriela (Éd.). 2017. *Lexicografia academică românească. Studii. Proiecte*: 221-234. Iași: Editura Universității "Alexandru Ioan Cuza".

Tamba, Elena Isabelle. 2017b. Informatizarea lexicografiei academice românești. Haja, Gabriela (Éd.). 2017. *Lexicografia academică românească. Studii. Proiecte*, Iași: Editura Universității "Alexandru Ioan Cuza": 145-155.

Tamba, Elena Isabelle. 2022a. CLRE. *Corpus lexicografic românesc electronic.* Începuturi, dezvoltare și perspective. Clim, Marius-Radu et Elena Isabelle Tamba (Coord.). 2022. *Actele Colocviului internațional "Lexicografia academică românească. Provocările informatizării", 2020–2021*: 219-232. Cluj-Napoca: Presa Universitară Clujeană.
URL: http://www.editura.ubbcluj.ro/bd/ebooks/pdf/3188.pdf

Tamba, Elena Isabelle. 2022b. The Role of the Institutes of the Romanian Academy in the Digitalization Process of the Linguistic Research. Jozić, Željko et Sabine Kirchmeier (Éds.). 2022. *The Role of National Language Institutions in the Digital Age. Contributions to the EFNIL Conference 2021 in Cavtat:* 91-100. Budapest: Nyelvtudományi Kukatóközpont.

Tamba Dănilă, E., M.-R. Clim, M. Patraşcu et A. Catană-Spenchiu. 2012. The Evolution of the Romanian Digitalized Lexicography. The Essential Romanian Lexicographic Corpus. Vatvedt Fjeld, R. et J.M. Torjusen (Éds.). 2012. *Proceedings of the 15th EURALEX International Congress, 7–11 August, 2012,* Oslo: 1014-1017. Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo.
URL: http://www.euralex.org/proceedings-toc/euralex_2012/

Trandabăţ, D., E. Irimia, V. Barbu Mititelu, D. Cristea et D. Tufiș. 2012. *Limba română în era digitală — The Romanian Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg/New York/Dordrecht/London: Springer.

Dictionnaires, corpus lexicographiques

CLRE: *Corpus lexicografic românesc electronic*, realizat de Departamentul de lexicologie și lexicografie, de la Institutul de Filologie Română "Alexandru Philippide" al Academiei Române — Filiala Iași. https://clre.solirom.ro/

DA: Academia Română, *Dicţionarul limbii române* [*DA*] (1913–1948), ediție digitală anastatică în *Corpus lexicografic românesc electronic*, realizat de Departamentul de lexicologie și lexicografie, de la Institutul de Filologie Română "Alexandru Philippide" al Academiei Române — Filiala Iași. https://dlr1.solirom.ro/

DAF: *Dictionnaire de l'Académie Française*. https://dictionnaire-academie.fr/

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1857 (Article)

DLR: Academia Română, *Dicţionarul limbii române* [*DLR*] (1965–2010), ediție digitală anastatică în *Corpus lexicografic românesc electronic*, realizat de Departamentul de lexicologie și lexicografie, de la Institutul de Filologie Română "Alexandru Philippide" al Academiei Române — Filiala Iași.
https://dlr1.solirom.ro/

DLR²: *Dicționarul limbii române* [*DLR*] (2021–présent) Ediția a doua, revizuită și adăugită. Academia Română. București: Editura Academiei Române.
https://dlri.ro/

DÉRom: Buchi, Eva et Wolfgang Schweickard (Dir.). *Dictionnaire Étymologique Roman* [*DERom]*, (2008–présent), Nancy, ATILF; publication éléctronique.
http://www.atilf.fr/DERom

DRAE: *Diccionario de la lengua española de la Real Academia Española.* http://buscon.rae.es/draeI/

DWB: *Deutsches Wörterbuch "der Grimm"*. http://germazope.uni-trier.de/Projects/DWB

Lesicon romanescu–latinescu–ungurescu–nemtescu quare de mai mulţi autori, in cursul a trideci, si mai multoru ani s'au lucrat. Seu Lexicon valachico–latino–hungarico–germanicum quod a pluribus auctoribus decursu triginta et amplius annorum elaboratum est. Budae, Typis et Sumtibus Typografiae Regiae Universitatis Hungaricae, 1825. Ediție electronică de Maria Aldea, Daniel-Corneliu Leucuţa, Lilla-Marta Vremir, Vasilica Eugenia Cristea și Adrian Aurel Podaru, Cluj-Napoca, 2013.
http://bcucluj.ro/lexiconuldelabuda/site/login.php

OED: *Oxford English Dictionary.* http://www.oed.com/

TLFi: *Le Trésor de la Langue Française Informatisé.* http://atilf.atilf.fr

TLIO: *Tesoro della lingua italiana delle origini*. http://tlio.ovi.cnr.it/TLIO/index2.html

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1862 (Article)

Training an AI-based Writing Assistant for Spanish Learners: The Usefulness of Chatbots and the Indispensability of Human-assisted Intelligence

Ángel Huete-García, *University of Oxford, Faculty of Medieval and Modern Languages, United Kingdom (angel.huete-garcia@mod-langs.ox.ac.uk) (https://orcid.org/0000-0002-8098-7401)*
and
Sven Tarp, *Aarhus University, Denmark, and Stellenbosch University, South Africa (st@cc.au.dk) (https://orcid.org/0000-0003-1941-9082)*

Abstract: This article deals with the relationship between human and artificial intelligence in the context of an ongoing Spanish Writing Assistant project, where ChatGPT is used to assist in four key tasks related to either training the underlying language model or preparing future user communication. The project is an interdisciplinary collaboration between lexicographers with experience in language teaching and IT experts from a high-tech company. The article first describes the methodology of the overall project and the specific role of the lexicographers. It then discusses the three tasks in which the latter are directly involved: the construction of a set of two parallel Spanish corpora, one correct and the other with induced errors, the generation of validation material, and the writing of extended grammatical explanations for Spanish learners. Based on a large amount of empirical data, including 35,000 carefully reviewed sentences, the article details the different steps of the interaction between human and chatbot, as well as the experiences and reflections drawn from this process. It concludes that the two parts engage in very different types of relationships depending on the concrete task, and that human knowledge, culture, skills and language intuition are crucial for the chatbot to work properly.

Keywords: SPANISH WRITING ASSISTANTS, LANGUAGE LEARNING, CHATBOTS, HUMAN-ASSISTED INTELLIGENCE, TRAINING OF LANGUAGE MODEL, CORPUS BUILDING

Opsomming: Die afrigting van 'n KI-gebaseerde skryfhulpmiddel vir Spaanse leerders: Die nut van kletsbotte en die onmisbaarheid van mensgesteunde intelligensie. Hierdie artikel handel oor die verhouding tussen menslike en kunsmatige intelligensie binne die konteks van 'n lopende Spaanse skryfhulpmiddelprojek, waar ChatGPT vir vier sleuteltake verwant aan óf die afrigting van die onderliggende taalmodel óf die voorbereiding van toekomstige gebruikerskommunikasie aangewend word. Dit is 'n interdissiplinêre samewerkingsprojek tussen leksikograwe met ervaring in taalonderrig en IT-spesialiste van'n hoëtegnologiemaatskappy. In die artikel word die metodologie van die algehele projek en die spesi-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1862 (Article)

fieke rol van die leksikograwe eers beskryf. Daarna word die drie take waarby die laasgenoemdes direk betrokke is, beskryf: die samestelling van 'n stel van twee parallelle Spaanse korpora, een korrek en die ander met ingevoerde foute, die generering van geldigheidsmateriaal, en die skryf van uitgebreide grammatikale verklarings vir Spaanse leerders. Gebaseer op 'n groot hoeveelheid empiriese data, insluitend 35,000 versigtig beoordeelde sinne, word die verskillende stappe van die interaksie tussen mens en kletsbot, sowel as die ervarings en refleksies wat uit die proses verkry is, uiteengesit. Daar word tot die gevolgtrekking gekom dat die twee dele, afhangende van die konkrete taak, in baie verskillende tipes verhoudings betrokke is, en dat menslike kennis, kultuur, vaardighede en taalintuïsie noodsaaklik is vir die kletsbot om behoorlik te funksioneer.

Sleutelwoorde: SPAANSE SKRYFHULPMIDDELS, TAALAANLEER, KLETSBOTTE, MENSGESTEUNDE INTELLIGENSIE, TAALMODELAFRIGTING, KORPUSBOU

1. Introduction

The term *artificial intelligence* was coined in 1955 by the computer scientist John McCarthy and three of his colleagues in a proposal they drafted for a summer research project at Dartmouth College the following year; see McCarthy et al. (2006). The idea was to study and progressively understand intelligence by implementing its essential features in artificial hardware (computers), rather than using natural biological cells and tissues for this purpose. As Noam Chomsky, who was closely involved in the discussions from the beginning, has argued, this deviated from the original goal shared by colleagues in other disciplines who were more interested in knowing how intelligence actually works in humans and is encoded in their genes, an interest that soon materialised in what is now known as *cognitive science*; see Katz (2012).

Be that as it may, artificial intelligence has undergone a remarkable development in recent years, proving itself capable of processing large amounts of data in an astonishingly short time that humans cannot match, and even generating new data from this material. In this context, various AI-based chatbots have been introduced and tested as teaching tools in language education, as discussed by Coniam (2008) and Yang et al. (2022), among others.

Particularly since the introduction of ChatGPT in late 2022, the practical use of artificial intelligence has taken something of a quantum leap, becoming almost commonplace among many scholars in a wide range of disciplines, including linguistics, language education and lexicography. It is now generally accepted that artificial intelligence presents both opportunities and risks, the severity of which obviously depends on the nature of each discipline. The same Chomsky (2023), for example, sees the use of ChatGPT as "basically high-tech plagiarism" and "a way of avoiding learning". However, he concedes that this and similar chatbots "may have some value for something", although "it's not obvious what".

This article reports on part of our experience from an ongoing Spanish Writing Assistant project where the use of ChatGPT is clearly helpful, as it significantly increases productivity without being perfect. Here, ChatGPT's generative AI-based language model is used to train and prepare another AI-driven language model called GECToR (*Grammatical Error Correction: Tag, Not Rewrite*), which will serve as a support for the writing assistant, just as it serves other similar tools such as *Grammarly*, as explained by Omelianchuk et al. (2020).

In this context, we have some reservations about the convenience of using the very term *artificial intelligence*. While the adjective *artificial* modifies the noun *intelligence* and contrasts it with *natural* intelligence, the noun itself still carries some false connotations and may cause confusion. We have had the same kind of conversations with ChatGPT as Chomsky et al. (2023), where it openly admits that it has serious limitations, now and for a long time to come, in terms of dreaming, reflecting and reasoning like humans, limitations that will have a noticeable impact on its performance and will require some degree of human intervention. Nevertheless, and despite our reservations, we will continue to use the term *artificial intelligence* because it is far too well established to change. For the sake of balance, however, we have used the term *human-assisted intelligence* in this article and its title with the same meaning as Huang and Tarp (2021), i.e. to complement artificial intelligence with real human intelligence when the former fails and proves insufficient.

In the following sections, we will discuss some of our experiences with using ChatGPT to solve specific types of tasks, and show how humans and artificial intelligence engage in different types of relationships depending on the concrete task at hand. Section 2 describes the ongoing Spanish Writing Assistant project, both the original plan and how it was modified in early 2023 after the introduction of ChatGPT. Section 3 then discusses the results and lessons learned from generating a set of two parallel Spanish corpora, one correct and one incorrect, for internal training purposes. Section 4 does something similar, based on the experience of generating internal validation data. Section 5 develops an idea that arose when the chatbot spontaneously started explaining certain grammatical issues without being asked, and shows some of its implications for external communication between the writing assistant and its users. Finally, Section 6 summarises the experience so far and draws some initial conclusions that will hopefully inspire other researchers.

2. Description of the project

The aim of the ongoing project is to develop an AI-based writing assistant for Spanish learners, both native and non-native. The project is an interdisciplinary collaboration between computer specialists from the Danish company *Ordbogen A/S* and a small team of lexicographers from Spanish, British and Danish universities. As Tarp and Gouws (2023) have argued, it seems quite natural that lexicographers with their user-oriented tradition should be involved in this kind of work. Indeed, many of them have already been engaged in various writing assistant projects in recent years; see Verlinde (2011), Granger and Paquot (2015), Tarp et al. (2017), Alonso-Ramos and García-Salido (2019), Frankenberg-García et al. (2019), Tarp (2020), Frankenberg-García (2020) and Fuertes-Olivera and Tarp (2020), among others. However, many of these projects are now being overtaken by AI-based technological developments that require significant investment and computational power.

The three main researchers involved in the current project, two of whom are native Spanish speakers, have — apart from traditional lexicographical competence — more than 60 years of combined experience in teaching Spanish to native and non-native students, the latter mainly with English, Chinese, Italian or Danish as their mother tongue. Against this backdrop, and unlike monolingual and mostly English AI-based writing assistants such as *DeepL Write*, *Grammarly*, *Ginger*, *LanguageTool* and *ProWritingAid*, the writing tool under construction will have, besides a fully Spanish version, also bilingual versions with comments and, when required, supplementary explanations written in the target users' native language, as well as various types of wake-up calls when particular challenges show up. The new tool, which will share some features with existing writing assistants like *Grammarly*, will also differ from them in that it will have a didactic function focused on Spanish language learning, in addition to helping learners write Spanish texts. This function will be expressed not only in the use of the learner's mother tongue, but above all in the straightforward explanation of orthographic, grammatical and, to some extent, semantic errors or confusions that learners often make.

The motivation for this design is that at the same time as writing in both native and non-native languages is increasingly, and sometimes exclusively, done on laptops, tablets and smartphones, written language is deteriorating in many places, especially among young people; see Carter and Harper (2013). This calls for new didactic methods that can motivate learners in new ways, so it seems logical to start where people write, on the devices mentioned above. Instead of being passive writing tools, these devices can be transformed — by incorporating writing assistants of the type outlined — into active tools that interact with users and their written language in different ways.

The project was drafted with a detailed work plan in the third trimester of 2022. Financial support was granted in December of the same year, and the project as such started in the early months of 2023. It is somehow related to three monolingual (Danish, German and English) writing assistant projects at *Ordbogen A/S*, but differs in important aspects in that it was formulated as a research project rather than a commercial one. This difference is mainly expressed in its didactic function and bilingual dimensions, but also in the experimental way it is carried out.

The initial work plan, as reproduced by Tarp (2023), was modified several times as the computer specialists, who are also going through a learning process, introduced new techniques and suggested new methods from time to time, a normal adaptation practice when working with cutting-edge technol-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1862 (Article)

ogy. But the biggest change came in March 2023, when it was decided to experiment with ChatGPT to see which tasks it could help with in terms of productivity, without compromising the quality of the work done. This led not only to changes in the way we worked on some of the predicted tasks, but also to the formulation of two entirely new types of tasks. In this context, our main focus as researchers became the new types of relationships we established with the chatbot, i.e. between human and machine, which is also the main topic of this article.

From a research perspective, the overall project plan consists of four partially intertwined phases, of which the first two are relevant for this article:

1. training the AI-based language model, for which the GECToR model was chosen;
2. preparing good user communication, inspired by the ideas of Norman (2013);
3. testing on real users, mainly using qualitative methods;
4. publishing the findings and conclusions.

The first phase consists of four main tasks, the first of which is to train the GECToR language model on an existing corpus. This is done by splitting the corpus into its multiple sentences and automatically introducing between one and five errors in each of these sentences in order to teach the language model to distinguish between right and wrong. The model is then fed with so-called synthetic data from a lexicographical database, i.e. all the words and their inflected full forms contained in the database, together with their respective grammatical categories (part of speech, gender, number, person, tense and mood). The purpose of this is both to enable the model to recognise existing words and word forms, and to provide it with an internal language to communicate with the lexicographers when they start writing comments and explanations. These two tasks were carried out by the computer specialists, as was the third task, which was an innovation compared to the original plan. In this case, ChatGPT was asked to create a special corpus of texts on topics typical of the target group and written in a style similar to theirs. In this way, and using special techniques, a corpus of one million words was created overnight and then used to train the language models as described above, thereby demonstrating a very time-saving way of creating specific corpora for internal use only.

The fourth task, the creation of two parallel corpora, also for training purposes, is described in detail in Section 3. It represents an idea that arose with the launch of ChatGPT, and is another deviation from the original plan, but this time carried out by the lexicographers. In this way, the GECToR language model will be trained on three different corpora created with different methods and techniques, a procedure that is expected to increase the quality of the final product. The last task to be completed in this first phase of the project is the preparation of validation data to evaluate the performance of the language model on different parameters. This is also done by the lexicographers with ChatGPT, as will be explained in Section 4.

In the second phase of the current project, three main tasks have to be solved:

1. writing small comments or glosses in Spanish to explain both the problems detected by the writing assistant and the suggestions it makes;
2. writing additional explanations to provide more detailed information on vocabulary, grammar and style for didactic purposes;
3. automatically translating all these texts into English, Danish, Italian and Chinese, using the experience from another project at *Ordbogen A/S*; see Tarp (2022b).

In the future, it is planned to conduct experiments to see if it is possible and beneficial to use ChatGPT to solve the first of these three tasks. Until then, the plan is to use it only to support the second task, as described in Section 5.

From the above, it can be seen that there are four key tasks where ChatGPT is now being used to develop the Spanish writing assistant. The three that directly involve lexicographers will be discussed in the following sections.

3. Creating two parallel Spanish corpora

The construction of two parallel Spanish corpora is an innovation in several respects. Unlike traditional parallel corpora, which are bilingual or multilingual, this new set of corpora is monolingual and consists of two identical corpora, the only difference being that one represents correct Spanish and the other contains some deliberately induced errors. It also differs from the other two corpora used in this project, where the errors were introduced using special software. Here they are created by ChatGPT following the instructions of the lexicographers, with the incorrect corpus preceding the corrected one.

As complete novices, we had to go through a process of experimentation to learn how to interact with ChatGPT and instruct it to give us the kind of texts we wanted. In this way, we came to a similar conclusion as Panday-Shukla (2023), i.e. to write specific, clear, concise and contextualised prompts. However, we also had to find an easy way around the chatbot's built-in resistance to introducing errors into the generated texts, a resistance we cracked by telling it that we needed its help for didactic purposes. Another challenge was to achieve diversity, not only lexically, but also in terms of the types of errors produced. It was therefore decided to ask ChatGPT to write texts or essays on 40 different topics, also selected with its help, and at the same time instruct it to play the role of a Spanish learner.

The most challenging aspect, however, has been that in order to train the GECToR language model appropriately, it is necessary to distinguish between two different classes of errors: *misspellings* and *word confusions*. The latter are errors that mix up existing words or word forms, most of which have almost identical pronunciations or spellings, such as *bello* vs. *vello* (*beautiful* vs. *hair*), or *bienes* vs. *vienes* (*assets* vs. *you come*). The problem here is that there is a grey area between the two classes of error. For example, if a learner spells *bello* with a *v* in *El vestido de gala es bello* (*The ball gown is beautiful*), this would normally be considered a spelling mistake in the classroom, but in the GECToR model it is a word confusion because both the adjective *bello* and the noun *vello* are common Spanish words that are part of the synthetic data also used to train the model (see Section 2).

As will be seen in the following subsections, it proved relatively easy to instruct ChatGPT to produce texts with the required types of misspellings, whereas it remains a challenge to achieve the required types of word confusions without writing specific prompts for each specific pair of confused words.

3.1 Generating texts with orthographic mistakes

In this subsection we will discuss the strategies used to make ChatGPT generate texts with two classes of orthographic mistakes typical of Spanish learners, those that are common to most learners and those that could be made by learners with a language disorder. In both cases, the incorrect texts are accompanied by the corrected versions. The prompts that contain our instructions to the chatbot were originally written in Spanish, but have been translated into English for the benefit of the reader.

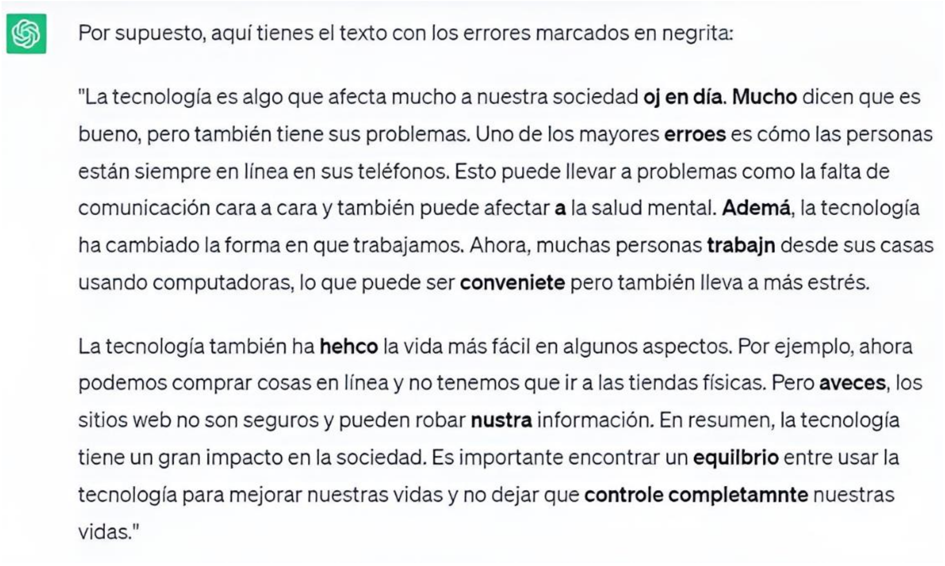
A lot of testing was done to find the best way to get the types of texts needed. Based on the tests, we developed a first master prompt that specified the objective, the role ChatGPT should play, and the task to be performed, i.e. to write an essay on one of the forty pre-selected topics with a series of common spelling mistakes. These topics can then be used in turn to create a total of 40 specific prompts. It should be noted that before presenting a prompt to the chatbot, it is necessary to give it the following introductory text to contextualise the topic:

> I am training a linguistic model for didactic purposes and I need you to take on the role of a Spanish student to write some texts.

As for the master prompt into which the pre-selected topics can be inserted, the following instructions seem to be the best to help ChatGPT understand the context in which it must operate:

> Write an essay on [the importance of protecting the environment] with real spelling mistakes, as if you were a real Spanish student. Then write the same text with the mistakes corrected.

In this master prompt, ChatGPT is asked to generate texts with unspecified spelling mistakes, giving it complete freedom to simulate these mistakes. The task of the lexicographers is then to check that the 40 specific prompts produce the requested data without any unforeseen problems. An example of a trial with general misspellings produced in response to this type of prompt is shown in Figure 1.



Por supuesto, aquí tienes el texto con los errores marcados en negrita:

"La tecnología es algo que afecta mucho a nuestra sociedad **oj en día**. **Mucho** dicen que es bueno, pero también tiene sus problemas. Uno de los mayores **erroes** es cómo las personas están siempre en línea en sus teléfonos. Esto puede llevar a problemas como la falta de comunicación cara a cara y también puede afectar **a** la salud mental. **Ademá**, la tecnología ha cambiado la forma en que trabajamos. Ahora, muchas personas **trabajn** desde sus casas usando computadoras, lo que puede ser **conveniete** pero también lleva a más estrés.

La tecnología también ha **hehco** la vida más fácil en algunos aspectos. Por ejemplo, ahora podemos comprar cosas en línea y no tenemos que ir a las tiendas físicas. Pero **aveces**, los sitios web no son seguros y pueden robar **nustra** información. En resumen, la tecnología tiene un gran impacto en la sociedad. Es importante encontrar un **equilbrio** entre usar la tecnología para mejorar nuestras vidas y no dejar que **controle completamnte** nuestras vidas."

Figure 1: Text with general spellings mistakes generated by ChatGPT

It is worth noting that the role type is a key element in the construction of the described set of parallel corpora, as it makes the chatbot assume a role that is as realistic as possible. The results show that it can relatively easily simulate real spelling mistakes, in addition to providing the corrected version for didactic purposes. Without specifying the role, context and didactic function, the chatbot may not be able to produce data with the requested misspellings or generate them automatically. However, even with this method, the mistakes may end up being repetitive and lacking variety. ChatGPT may also occasionally mark correctly spelled words as misspellings, but this is not an issue as the other generated mistakes are sufficient (see Figure 1).

One of the great advantages of asking the chatbot to imitate a student is that it also opens up a more complex and specific dimension in the generation of misspellings for the set of parallel corpora. It is known that some students suffer from a Specific Learning Disorder (SLD) related to the writing process, the *dysorthographia*, which affects their performance and, consequently, their learning; see Chung et al. (2020). We have explored this approach in order to make the types of spelling mistakes included in our parallel corpora more varied and specific. In this sense, the ability of ChatGPT to simulate psycholinguistic disorders such as dysorthographia is an interesting discovery that allows the inclusion of these phenomena in parallel corpora. To do this, it is necessary to specify the type of language disorder that the chatbot should reproduce. However, as ChatGPT confused *dysorthographia* with *stuttering*, the term was replaced by the synonymous *linguistic dysgraphia* (Chung et al. 2020, S47), resulting in a second type of master prompt:

> Write an essay on [the challenges of teaching and learning] using spelling mistakes frequently made by a Spanish learner with linguistic dysgraphia. Then write the same text with the mistakes corrected.

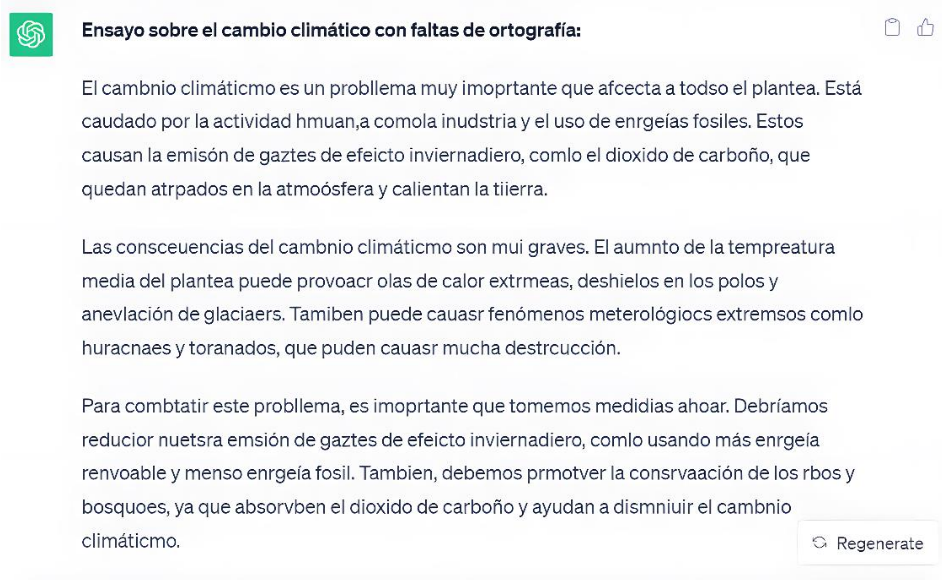
Searching with this prompt allows the corpus of errors to be expanded to include some of the less common errors that are associated with language disorders. The examples extracted in this way will undoubtedly help the GECToR model to respond to more specific problems that have not received specific attention from other writing assistants. The results of the first tests to emulate dysorthographic processes were favourable, with the generation of texts containing lexical errors such as syllable changes and confusion of phonemes or orthographic rules:

— Laglobalisazión* = La globalización
— puevlos* = pueblos
— homegenización* = homogeneización
— privacida* = privacidad
— conscienca* = consciencia

or complete sequences such as:

— Para abordar esto, es inportante eduvar a las persnas sobr los riesgos* = Para abordar esto, es importante educar a las personas sobre los riesgos.

Comparing the texts in Figures 1 and 2, it is obvious that the texts generated with this last approach contain more spelling mistakes than when the chatbot is asked to produce misspellings in general, although for our corpora it is not so much the quantity as the variety that is important. Moreover, due to the nature of the disorder itself, the errors in this case are orthographic, as can also be seen in Figure 2. In any case, the lexicographer will still need to quickly assess the errors to see if the 40 specific prompts related to the selected topics are appropriate for the purpose, so that the chatbot does not provide unwanted data, before handing them over to the computer specialists for automated serial production and creation of two parallel corpora of the required size.



Ensayo sobre el cambio climático con faltas de ortografía:

El cambnio climáticmo es un probllema muy imoprtante que afcecta a todso el plantea. Está caudado por la actividad hmuan,a comola inudstria y el uso de enrgeías fosiles. Estos causan la emisón de gaztes de efeicto inviernadiero, comlo el dioxido de carboño, que quedan atrpados en la atmoósfera y calientan la tiierra.

Las consceuencias del cambnio climáticmo son mui graves. El aumnto de la tempreatura media del plantea puede provoacr olas de calor extrmeas, deshielos en los polos y anevlación de glaciaers. Tamiben puede cauasr fenómenos meterológiocs extremsos comlo huracnaes y toranados, que puden cauasr mucha destrcucción.

Para combtatir este probllema, es imoprtante que tomemos medidias ahoar. Debríamos reducior nuetsra emsión de gaztes de efeicto inviernadiero, comlo usando más enrgeía renvoable y menso enrgeía fosil. Tambien, debemos prmotver la consrvaación de los rbos y bosquoes, ya que absorvben el dioxido de carboño y ayudan a dismniuir el cambnio climáticmo.

Figure 2: Text with specific dysorthographic mistakes generated by ChatGPT

3.2 Generating texts with word confusions

As mentioned above, generating word confusions with ChatGPT is much more complicated than generating misspellings. The list of the former is quite long because it also includes the use of acute accents (see the discussion of *-ar* verbs in Section 4). So far, we have not found a way to design a master prompt that meets this challenge without indicating the specific word pair that is being confused. Until we find the philosopher's stone, we are therefore working with specific prompts that specify both the topic and the confused word pair in question. We have started here with the most common problems related to the word pairs:

— *ser* and *estar* (both meaning *to be*);
— with and without an acute accent like *mi* vs. *mí* (*my* vs. *me*);
— with and without a silent *h* like *hola* vs. *ola* (*hello* vs. *wave*);
— spelled with *b* and *v* like *tubo* vs. *tuvo* (*pipe* vs. *she had*);
— spelled with *c* and *s* like *cien* vs. *sien* (*hundred* vs. *temple*);
— spelled with *s* and *z* like *casa* vs. *caza* (*house* vs. *hunting*);
— spelled with *ll* and *y* like *valla* vs. *vaya* (*fence* vs. *damn*);
— etc.

To create the concrete list, we drew on a similar list of word confusions that was compiled for the purpose of generating validation data (see Section 4). The

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1862 (Article)

specific prompts used to elicit a particular type of error in the texts require a precise description of that type of error, as shown here:

> Take the role of a Spanish student and write an essay on [the importance of protecting the environment] in which [the verbs 'ser' and 'estar'] are confused. Then write the same essay with the errors corrected.

The critical element here is the confusion of the verbs *ser* and *estar*. As this is very concrete, ChatGPT easily generates the corresponding word confusions. The next goal is to increase the variety by finding the formula to induce this type of error on a more general level, without having to specify each problem separately.

In this whole process, the role of the lexicographer is essential to obtain the desired text samples with ChatGPT, as well as to monitor the material provided, without the need for in-depth reading. The objective is to ensure that the prompt works properly and can be passed on for automatic use in the compilation of the required corpora, which are only intended for internal training and therefore do not need further revision, as minor errors are irrelevant.

With the methodology presented in this section, it is possible to build a large set of parallel corpora in a relatively short time, which would otherwise take longer and require licences to build from authentic texts.

4. Generating validation material

Once the GECToR language model has been trained on the three types of corpora mentioned above, it is necessary to have a tool for measuring its performance in order to determine when it has reached a satisfactory level that makes it suitable for real user testing. If it is underperforming, further training on a larger set of corpora will be required. For this purpose, it is necessary to generate validation data based on at least 100 of the most common errors in written Spanish, either spelling mistakes or confusion of existing words or inflectional forms. Then, for each of these types of errors, a set of parallel sentences will be created with these errors and their corrections. Unlike the corpora, the sentences must be 100 percent as required — both the correct ones without errors and the incorrect ones with their specific errors — in order to obtain a reliable measurement tool. This requires meticulous proofreading and editing by skilled lexicographers with a good command of Spanish. (For more details on this particular task, see Tarp and Nomdedeu-Rull 2024).

All this proved to be a learning process for both the computer specialists and the lexicographers. Initially, the task set by the former was to write 30 correct and 30 incorrect sentences of each type, of which 20 were used for training and 10 for validation, but these numbers were gradually increased. After the implementation of ChatGPT in the project, with the possibility of generating parallel corpora as described in the previous section, the number was set at 100 correct and 100 incorrect sentences for validation purposes only. For common misspellings, where the misspelled word does not exist, this means 100 sentences with the correct word and another 100 with the misspelling. By contrast, for word confusions, where both words exist in Spanish, two times 50 correct sentences are required, where each word is given in its correct context, together with two times 50 incorrect sentences, so that the language model is not misled into "thinking" that one of them does not exist.

The first challenge for the lexicographers was to identify the most common errors in written Spanish. We could not find a comprehensive inventory anywhere, only a few smaller lists that were somewhat useful for our purposes. Fortunately, the Centre for Applied Linguistics in Santiago de Cuba sent us a list of recorded spelling mistakes made by Cuban schoolchildren (Ruiz-Miyares 2016), from which we selected the most frequent ones. But that was still not enough. So we decided to ask ChatGPT, and it did indeed come up with some suggestions that we considered useful, based on our accumulated teaching experience, but also other suggestions, some of which were not even real mistakes. Finally, we brainstormed, using our experience and linguistic knowledge to identify more relevant errors. With this combined approach, we were able to compile a list of 172 common errors in written Spanish, not all of which may be the most frequent, but which together form a solid body of validation data for the specific purpose.

As in the previous section, the task now was to learn how to instruct ChatGPT to generate the desired data. After some experimentation, we developed a model where we first briefly introduce the problem and ask the chatbot if it is aware of it. An example of this, translated into English, is *Many people mistakenly confuse "asar" with "azar". I assume you know this*. The chatbot then immediately responds with a short description of the problem. This description allows us to see if it actually understands the problem, which it does about half the time. In the above case, it correctly classifies *asar* as a verb *referring to cooking food directly exposed to fire or dry heat, such as barbecuing meat* and *azar* as *a noun used to refer to randomness, coincidence, or luck*.

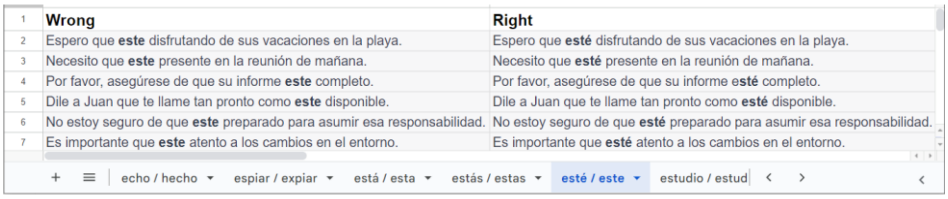
In the remaining cases, however, it either forgets to tell us that a particular word form can belong to more than one part of speech, or it simply gives incorrect examples of what it has correctly described at a more abstract level. For instance, after correctly explaining that the disruptive conjunction *o* (*or* in English) is written *u* when *it precedes words beginning with "o" or "ho", to avoid repeating the same sound in succession*, it immediately gives the following examples, the second of which is nonsensical:

— Tengo que elegir entre trabajar "o" estudiar (Correct)
— Tengo que elegir entre trabajar "u" estudiar (Correct, to avoid repeating the sound "o")

In such cases we immediately correct the chatbot, which then apologises almost mechanically for the misrepresentation and provides a more appropriate explanation. But the very fact that it can correctly explain a grammatical problem while at the same time giving incorrect examples of exactly the same problem made us suspect from the start that something was not quite working as it should, a suspicion that later turned out to be justified (see below).

When we are convinced, usually within a few seconds, that ChatGPT has at least partially grasped a problem, we tell it that we need its help to train a language model with a didactic purpose. This reference to language teaching turns out to be an effective way of encouraging it to also write incorrect sentences, which it has repeatedly refused to do, with the excuse that it is not allowed to do so for OpenAI.

After these initial manoeuvres, we tell the chatbot what we want it to do. Here we use a model of short, clear instructions, as recommended by Panday-Shukla (2023). Since ChatGPT can only generate a certain number of words at a time, we usually ask it to write 25–30 correct sentences, followed by 25–30 incorrect ones, all of which are copied into Google Sheets, where they are immediately reviewed and useless pairs of sentences are deleted (see Figure 3). The prompt is then repeated with some modifications to get linguistic variation without too many similar and stereotypical examples. This process continues until at least 100 valid sentence pairs have been generated.



1	**Wrong**	**Right**
2	Espero que **este** disfrutando de sus vacaciones en la playa.	Espero que **esté** disfrutando de sus vacaciones en la playa.
3	Necesito que **este** presente en la reunión de mañana.	Necesito que **esté** presente en la reunión de mañana.
4	Por favor, asegúrese de que su informe **este** completo.	Por favor, asegúrese de que su informe **esté** completo.
5	Dile a Juan que te llame tan pronto como **este** disponible.	Dile a Juan que te llame tan pronto como **esté** disponible.
6	No estoy seguro de que **este** preparado para asumir esa responsabilidad.	No estoy seguro de que **esté** preparado para asumir esa responsabilidad.
7	Es importante que **este** atento a los cambios en el entorno.	Es importante que **esté** atento a los cambios en el entorno.

Figure 3: Pair of correct and incorrect sentences generated by ChatGPT

For most problem types, the chatbot produces more than 95 usable sentence pairs out of 100 generated, which is an impressive performance. The remaining 3–5 pairs are problematic, either because it provides a different word or inflection than the one requested, because it forgets to replace the correct words with incorrect ones, or because it simply reformulates the sentence, making it useless for training purposes. Needless to say, it takes a well-trained human eye to spot these examples and sort them out. This is all the more true when, for one reason or another, the number of useless sentences grows and, in some cases, explode. The challenge seems to be greatest with word confusions, whereas the generation of sentence pairs containing only orthographic errors and their correction is relatively seamless.

An interesting case is *-ar* verbs, i.e. verbs that end in *-ar* in the infinitive and have a specific conjugation pattern. Many people, both native and non-native, often forget the acute accent on the *o* in the third person singular in the preterite of these verbs, e.g. *compró* (*she bought*), and write *compro* instead, which

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1862 (Article)

is the same form as the first-person singular in the present tense (*I buy*). This is usually considered a spelling mistake, but from the perspective of the language model it is a confusion of words, since both inflected forms exist. When ChatGPT was instructed to generate 50 correct sentences with each of these two inflected forms of *-ar* verbs and then a similar number of incorrect sentences, several challenges arose for what appears on the surface to be a homogeneous group of verbs.

Firstly, it proved essential to instruct the chatbot to create contexts that clearly indicate that it is either the first person singular present tense or the third person singular preterite tense. Otherwise the two inflected forms could fit into the same sentence, e.g. *tomo una cerveza* and *tomó una cerveza* (*I drink a beer* and *she drank a beer*), both of which are correct from a formal point of view, since Spanish personal pronouns such as *I* and *she* are often not used explicitly, but are implicitly expressed by the person-inflected verb forms. This requires not only grammatical knowledge to write precise instructions, but also linguistic intuition to proofread carefully and responsibly.

Secondly, a whole range of different problems arose for the individual verbs, not only because some of them have a similar noun form, such as *amo* (*I love* or *owner*) and *tomo* (*I take* or *volume*), which were used instead of the requested verb forms, but also because ChatGPT started inflecting the latter, thus providing irrelevant data. It was therefore necessary, to some extent, to give the chatbot tailored instructions to avoid too many useless examples. But even so, and even with a new string opened, it became increasingly obvious to us that there must be something inherent in these verbs that we are not yet aware of, probably related to their semantic and syntactic properties, that makes them behave differently and require special attention and differentiated treatment to be useful for our purpose.

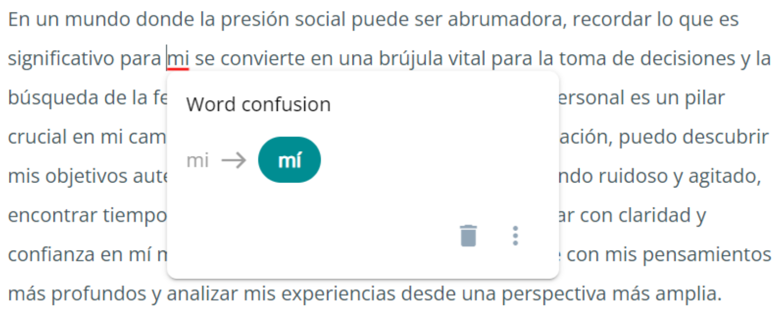
Thirdly, as with the *o* and *u* confusion above, the most worrying thing is that ChatGPT was asked several times along the way if it understood the actual problem. It claimed it did, with a correct explanation, and then continued to make mistakes. Such a wide gap between theory and practice reveals its inability to think and reason like a human being, as it admitted itself when questioned. It also demonstrates the need for assistance from genuine human intelligence.

For a few words, especially those with or without an acute accent, such as *mí* (*me*) and *mi* (*my*), where ChatGPT was struggling to generate errors, this supreme intelligence decided to cut the nonsense and instead to copy the correct sentences into a Word document and use the replace function to produce the errors in a much faster and more pragmatic way.

Thanks to the combined efforts of man and machine, the required validation material was generated in a surprisingly short time. Before ChatGPT, a skilled lexicographer could write 200 sentences a day in four hours before his or her brain ran out of energy. With ChatGPT, that figure is now 4,000 in the same time, with each set of correct and incorrect sentences taking anywhere from a few minutes to over half an hour to produce. This represents a 20-fold increase in productivity, proving the usefulness of chatbots despite their many shortcomings. In a matter of days, we built a corpus of around 35,000 correct and incorrect sentences to validate the performance of the language model. A literary critic might be sceptical about some of these sentences, but they are all formally correct and the problematic words are used in different combinations and contexts, making them suitable for the specific purpose.

5. Writing supplementary explanations

As explained in Section 2, the main objective of the ongoing project is to develop a Spanish writing assistant with a didactic function. This implies, among other things, providing future users with more detailed and easy-to-read explanations of particular types of orthographic, grammatical and semantic problems, such as the difference between *asar* and *azar*, between *mi* and *mí*, the use of *u* instead of *o*, etc.

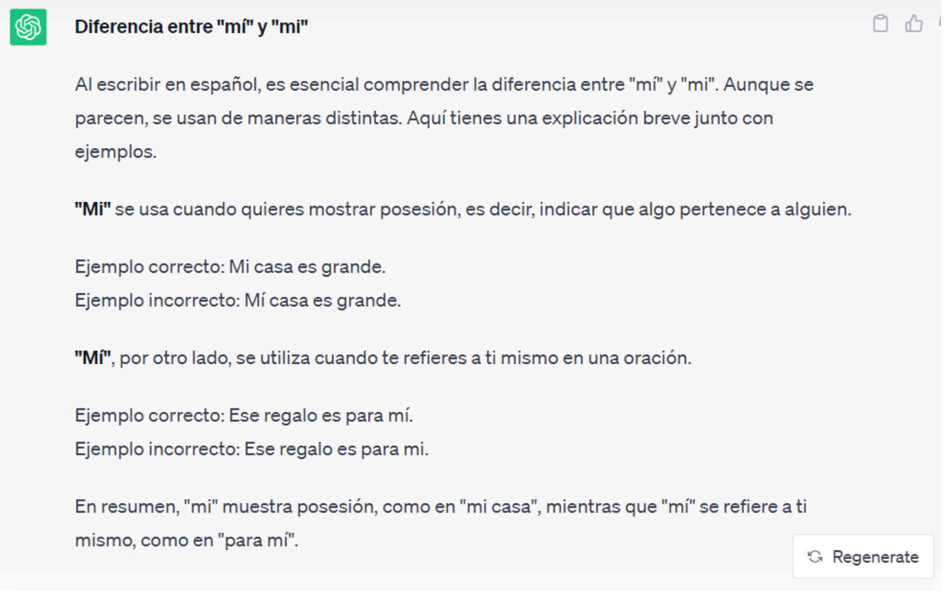


En un mundo donde la presión social puede ser abrumadora, recordar lo que es
significativo para mi se convierte en una brújula vital para la toma de decisiones y la
búsqueda de la fe ersonal es un pilar
crucial en mi cam ación, puedo descubrir
mis objetivos aut ndo ruidoso y agitado,
encontrar tiempo ar con claridad y
confianza en mí n con mis pensamientos
más profundos y analizar mis experiencias desde una perspectiva más amplia.

Figure 4: Writing assistant highlighting problem and suggesting alternative solution

Figure 4 shows how the writing assistant works after training the GECToR language model and before adding explanations and other didactic features. When users type something that the model detects as problematic, in this case *mi*, it is automatically underlined and they can then simply click on it to activate a pop-up window with an alternative suggestion (*mí*). The idea is to include, by default, a small comment that very briefly explains the problem and the suggestion without interrupting the writing flow, thus allowing for incidental learning as defined by Tarp (2022a). And in cases like the one in Figure 4, learners will also have the option of accessing a supplementary explanation that supports intentional learning.

The explanations required are fundamentally different from the traditional lexicographical definitions that explain the meaning of *lemmas*, in that here they have to explain *classes of problems*, i.e. grammatical, semantic and orthographic problems highlighted by the writing assistant, thus facilitating the learning of grammatical rules, word meanings and spelling. As Spanish teachers with decades of combined experience, we obviously know what these problems are and could easily explain them to our students in class. However, experience also tells us that it is less straightforward to write a concise explanation that gets to the heart of the matter in a language that is easily understood by the target audience. In addition to selecting the key aspects to be covered, determining the most appropriate and pedagogical structure can be quite time-consuming. We therefore decided to take inspiration from ChatGPT, which had already demonstrated its ability to write such extended explanations in the previous two tasks. In fact, we became aware of this ability when, without being prompted, it spontaneously began to explain grammatical problems related to these tasks. We therefore instructed it to write short didactic texts explaining the different problems identified to a student writing an essay in Spanish, giving both correct and incorrect examples.



Diferencia entre "mí" y "mi"

Al escribir en español, es esencial comprender la diferencia entre "mí" y "mi". Aunque se parecen, se usan de maneras distintas. Aquí tienes una explicación breve junto con ejemplos.

"Mi" se usa cuando quieres mostrar posesión, es decir, indicar que algo pertenece a alguien.

Ejemplo correcto: Mi casa es grande.
Ejemplo incorrecto: Mí casa es grande.

"Mí", por otro lado, se utiliza cuando te refieres a ti mismo en una oración.

Ejemplo correcto: Ese regalo es para mí.
Ejemplo incorrecto: Ese regalo es para mi.

En resumen, "mi" muestra posesión, como en "mi casa", mientras que "mí" se refiere a ti mismo, como en "para mí".

Figure 5: Supplementary explanation suggested by ChatGPT

Figure 5 shows a typical result on the difference between *mí* and *mi*. It appears with a reasonable overall structure, including a general introduction to the prob-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1862 (Article)

lem, an explanation of each word with correct and incorrect examples, and finally a short summary. However, it clearly needs some editing. For example, it might be helpful for some users to know that *mí* and *mi* are a pronoun and a possessive adjective respectively. The style should also be standardised, either by addressing the reader directly throughout the text or by not doing so at all. And there are a few other minor issues. For other types of problems, such as the confusion between the two past tenses in Spanish, it has also proved necessary to add more context to the example sentences in order to enhance their didactic role.

If the lexicographers are happy with an explanation suggested by the chatbot, they can make the appropriate changes or rewrite it according to the outlined structure. If they are not satisfied, or if they need further inspiration, they can simply click the "regenerate" button. The chatbot will then come back with another, slightly different suggestion, which may include some of the things mentioned above.

Experience has shown that very few suggestions can be used as they are. In most cases it is sufficient to make a few changes based on the lexicographers' knowledge and teaching experience, but experience has also shown that it can be very inspiring to get more suggestions. The creation of supplementary explanations is another fascinating experience of this project and shows the perspectives of the described forms of interaction between artificial and human intelligence.

6. Conclusion

In this article we have discussed three ways of using ChatGPT to develop a writing assistant for Spanish learners, either by training the underlying GECToR language model or by preparing external communication with users. In all three cases, the lexicographers had to be both open-minded and creative in their engagement with the chatbot in order to work out the instructions that would make it generate the required types of text. These examples represent three different types of relationships between the lexicographers and the chatbot, between man and machine:

1. When building parallel corpora for training, the lexicographers have to check that the prompts make the chatbot produce the right text types in the trials, but there is no need for them to proofread the texts that are later mass-generated and included in the corpora, as these are only for internal training purposes.
2. When producing validation material, it is necessary for the lexicographers to carefully proofread all example sentences in order to find and correct any mistakes made by the chatbot.
3. When preparing supplementary explanations, the chatbot's role is only to inspire, while the lexicographers' role is to adapt the text to the user.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1862 (Article)

Throughout this article, we have also discussed and seen examples of how ChatGPT lacks key features usually associated with intelligence. If these features are ignored or underestimated, the use of generative AI chatbots can indeed be risky, as both an eminent scientist like Chomsky (2023) and an experienced data analyst like Southern (2023) argue. But if they are taken into account, chatbots can be extremely useful, as we have seen above. In our specific case, it has significantly increased productivity, sometimes by an impressive 20 times, which is also something that should not be ignored.

The key question is who is the master, the human or the chatbot. The former should avoid being reduced to an appendage of the machine. In a sense, the new reality requires even more knowledge, more general culture, more skills and, in our case, more linguistic intuition to be able to interact with the chatbot appropriately, to give it precise and guiding instructions, to check the content and form of the texts it generates, and to cut it off when it simply does not serve the desired purpose. Human-assisted intelligence is certainly a must when working with ChatGPT.

Acknowledgments

Special thanks are due to Programmer and Web Developer Henrik Hoffmann, *Ordbogen A/S*, Denmark, for technical support, and Dr. Leonel Ruiz Miyares, Center for Applied Linguistics, Santiago de Cuba, for providing empirical data.

Thanks are also due to Professor Antoni Nomdedeu Rull, Rovira i Virgili University, Spain, Professor Pedro Fuertes-Olivera, Valladolid University, Spain, and Professor Rufus H. Gouws, Stellenbosch University, South Africa, for their constructive comments.

Finally, we would also like to thank the Aarhus University Research Foundation for funding a six-month sabbatical to conduct this research project.

References

A. Digital tools

ChatGPT: https://chat.openai.com
DeepL Write: https://www.deepl.com/write
Ginger: https://www.gingersoftware.com
Grammarly: https://www.grammarly.com
LanguageTool: https://languagetool.org
ProWritingAid: https://prowritingaid.com

B. Literature

Alonso-Ramos, M. and M. García-Salido. 2019. Testing the Use of a Collocation Retrieval Tool

Without Prior Training by Learners of Spanish. *International Journal of Lexicography* 32(4): 480-497. https://doi.org/10.1093/ijl/ecz016

Carter, M.J. and H. Harper. 2013. Student Writing: Strategies to Reverse Ongoing Decline. *Academic Questions* 26(3): 285-295.
Student Writing: Strategies to Reverse Ongoing Decline by Heather Harper | NAS

Chomsky, N. 2023. Noam Chomsky on ChatGPT: It's "Basically High-Tech Plagiarism" and "a Way of Avoiding Learning". *Open Culture*, February 10, 2023.
Noam Chomsky on ChatGPT: It's "Basically High-Tech Plagiarism" and "a Way of Avoiding Learning" | Open Culture

Chomsky, N., I. Roberts and J. Watumull. 2023. The False Promise of ChatGPT. *The New York Times*, March 8, 2023.
Opinion | Noam Chomsky: The False Promise of ChatGPT - The New York Times (nytimes.com)

Chung, P.J., D.R. Patel and I. Nizami. 2020. Disorder of Written Expression and Dysgraphia: Definition, Diagnosis, and Management. *Translational Pediatrics* 9(S1): S46-S54.
http://dx.doi.org/10.21037/tp.2019.11.01

Coniam, D. 2008. Evaluating the Language Resources of Chatbots for their Potential in English as a Second Language. *ReCALL* 20(1): 98-116.
https://doi.org/10.1017/S0958344008000815

Frankenberg-García, A. 2020. Combining User Needs, Lexicographic Data and Digital Writing Environments. *Language Teaching* 53(1): 29-43.
https://doi.org/10.1017/S0261444818000277

Frankenberg-García, A., R. Lew, J.C. Roberts, G.P. Rees and N. Sharma. 2019. Developing a Writing Assistant to Help EAP Writers with Collocations in Real Time. *ReCALL* 31(1): 23-39.
https://doi.org/10.1017/S0958344018000150

Fuertes-Olivera, P.A. and S. Tarp. 2020. A Window to the Future: Proposal for a Lexicography-assisted Writing Assistant. *Lexicographica* 36: 257-286.
https://doi.org/10.1515/lex-2020-0014

Granger, S. and M. Paquot. 2015. Electronic Lexicography Goes Local: Design and Structures of a Needs-driven Online Academic Writing Aid. *Lexicographica* 31(1): 118-141.
https://doi.org/10.1515/lexi-2015-0007

Huang, F. and S. Tarp. 2021. Dictionaries Integrated into English Learning Apps: Critical Comments and Suggestions for Improvement. *Lexikos* 31(1): 68-92.
https://doi.org/10.5788/31-1-1626

Katz, Y. 2012. Noam Chomsky on Where Artificial Intelligence Went Wrong. An Extended Conversation with the Legendary Linguist. *The Atlantic*, November 1, 2012.
Noam Chomsky: Where Artificial Intelligence Went Wrong - The Atlantic

McCarthy, J., M.L. Minsky, N. Rochester and C.E. Shannon. 2006. A Proposal for the Dartmouth Summer Research Project on Artificial Intelligence, August 31, 1955. *AI Magazine* 27(4): 12.
https://doi.org/10.1609/aimag.v27i4.1904

Norman, D. 2013. *The Design of Everyday Things*. New York: Basic Books.
The Design of Everyday Things (archive.org)

Omelianchuk, K., V. Atrasevych, A. Chernodub and O. Skurzhanskyi. 2020. GECToR — Grammatical Error Correction: Tag, Not Rewrite. Burstein, J., E. Kochmar, C. Leacock, N. Madnani, I. Pilán, H. Yannakoudakis and T. Zesch (Eds.). 2020, *Proceedings of the 15th Workshop on Inno-*

vative Use of NLP for Building Educational Applications: 163-170. Seattle: Association for Computational Linguistics.
http://dx.doi.org/10.18653/v1/2020.bea-1.16

Panday-Shukla, P. 2023. Five Things to Know about Generative Artificial Intelligence. *Galico Infobytes*, June, 2023.
Infobyte_June-2023.pdf (calico.org)

Ruiz-Miyares, L. 2016. ¿Cómo está la ortografía en 6to, 9no y 12mo grados en Santiago de Cuba? *Revista Ciencias Pedagógicas* 9(3): 1-15.
https://www.cienciaspedagogicas.rimed.cu/index.php/ICCP

Southern, B. 2023. I've Worked as a Data Analyst at Companies like Amazon for 20 Years. Using ChatGPT for Data Analytics Is a Risky Move — AI Can't Do the Work We Do. *Business Insider*, July 20, 2023.
Here's Why Leaders Shouldn't Use ChatGPT for Data Analytics (businessinsider.com)

Tarp, S. 2020. Integrated Writing Assistants and their Possible Consequences for Foreign-Language Writing and Learning. Bocanegra-Valle, A. (Ed.). 2020. *Applied Linguistics and Knowledge Transfer: Employability, Internationalization and Social Challenges*: 53-76. Bern: Peter Lang.
https://doi.org/10.3726/b16992

Tarp, S. 2022a. A Lexicographical Perspective to Intentional and Incidental Learning: Approaching an Old Question from a New Angle. *Lexikos* 32(2): 203-222.
https://doi.org/10.5788/32-2-1703

Tarp, S. 2022b. Turning Bilingual Lexicography Upside Down: Improving Quality and Productivity with New Methods and Technology. *Lexikos* 32: 66-87.
https://doi.org/10.5788/32-1-1686

Tarp, S. 2023. Eppur si muove: Lexicography Is Becoming Intelligent! *Lexikos* 33(2): 107-131.
https://doi.org/10.5788/33-2-1841

Tarp, S., K. Fisker and P. Sepstrup. 2017. L2 Writing Assistants and Context-Aware Dictionaries: New Challenges to Lexicography. *Lexikos* 27: 494-521.
http://dx.doi.org/10.5788/27-1-1412

Tarp, S. and R.H. Gouws. 2023. A Necessary Redefinition of Lexicography in the Digital Age: Glossography, Dictionography and the Implications for the Future. *Lexikos* 33: 425-447.
https://doi.org/10.5788/33-1-1826

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* 97: 309-321.
https://dx.doi.org/10.5209/clac.91985

Verlinde, S. 2011. Modelling Interactive Reading, Translation and Writing Assistants. Fuertes-Olivera, P.A. and H. Bergenholtz (Eds.). 2011. *e-Lexicography: The Internet, Digital Initiatives and Lexicography*: 275-286. London/New York: Continuum.
https://doi.org/10.5040/9781474211833.ch-013

Yang, H., H. Kim, J.H. Lee and D. Shin. 2022. Implementation of an AI Chatbot as an English Conversation Partner in EFL Speaking Classes. *ReCALL* 34(3): 327-343.
https://doi.org/10.1017/S0958344022000039

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1871 (Lexiconote)

On the Inclusion of Neologisms in *Oxford Advanced Learner's Dictionary* (10th edition)

Anmin Wang, *School of Foreign Studies, Guangxi Minzu University, Nanning, China (anmin.wang@gxmzu.edu.cn) (https://orcid.org/0009-0005-2132-9539)*
and
Xi Chen, *School of Foreign Studies, Guangxi Minzu University, Nanning, China* (*cymter@live.com) (https://orcid.org/0009-0006-3231-0830)*

Abstract: Adding neologisms to a dictionary in its revision helps keep it abreast of time, which applies to a learner's one like *Oxford Advanced Learner's Dictionary* (10th edition) (henceforth OALD 10), its latest edition. English Learner's dictionaries like OALD 10 get revised regularly. In this article, the neologisms included in OALD 10 have been approached from different perspectives. In terms of the part of speech, 71.1% of the neologisms are nouns, which could speak for the nouny nature of English. Content words like nouns, adjectives, adverbs and verbs make up almost 96% of the neologisms. Judging by the word-formation, the top three ways to form the neologisms are compounding, derivation and blending in decreasing order. According to the form the compounds take, which are mostly nouns, the open, hyphenated and tight ones come in decreasing order. 623 compounds make up more than half of the neologisms, which provides strong evidence for compounding to be the most frequently used way of creating neologisms. For neologisms formed by derivation, most of them are also nouns. However, 170 out of 300 neologisms formed by derivation have already appeared as derivatives of headwords in OALD 9. For 34 blends, 75% of them are partial in nature, which means at least one word in making a blend is in its full form. When it comes to the new words with regional labels, nearly 80% are labeled as belonging to British and North American usages, a sign of the hidden Anglo-centrism. The status of some words counted as neologisms is questionable, as they have been in use for a very long time. The impact of science and technology on the inclusion of neologisms in OALD 10 is quite visible. There are still many opportunities for further exploration concerning OALD 10.

Keywords: NEOLOGISM, OALD 10, PART-OF-SPEECH, WORD-FORMATION, REGIONAL LABEL, ANGLO-CENTRISM, DICTIONARY REVISION, SCI-TECH INFLUENCE

Opsomming: Oor die insluiting van neologismes in die *Oxford Advanced Learner's Dictionary* (10de uitgawe). Om neologismes by te voeg by 'n woordeboek in die hersiening daarvan, help om die woordeboek bygewerk te hou, wat relevant is vir 'n aanleerderswoordeboek soos *Oxford Advanced Learner's Dictionary* (10de uitgawe) (voortaan OALD 10), die jongste uitgawe. Engelse aanleerderswoordeboeke soos OALD 10 word gereeld hersien. In hierdie artikel word die neologismes wat in OALD 10 ingesluit is, vanuit verskillende perspektiewe benader. Betreffende die woordsoort is 71.1% van die neologismes selfstandige naamwoorde, wat die naam-

Lexikos 34 (AFRILEX-reeks/series 34: 2024): 41-50

woordelike aard van Engels bevestig. Inhoudswoorde soos selfstandige naamwoorde, adjektiewe, bywoorde en werkwoorde beslaan ongeveer 96% van die neologismes. Te oordeel na woordvorming is, in dalende volgorde, die drie belangrikste metodes waarmee neologismes gevorm word samestelling, afleiding en samesmelting. Volgens die vorm wat die samestellings, wat meestal selfstandige naamwoorde is, aanneem, is hulle in dalende volgorde oop, koppelteken- en geslote samestellings. 623 samestellings vorm meer as die helfte van die neologismes wat kragtige bewys bied dat samestelling die mees frekwente metode is wat gebruik word om neologismes te vorm. Betreffende neologismes wat deur afleiding gevorm is, is die meeste ook selfstandige naamwoorde. 170 uit 300 neologismes wat deur afleiding gevorm is, het reeds as afleidings van trefwoorde in OALD 9 verskyn. Rakende 34 samesmeltings is 75% van hulle gedeeltelik van aard, wat daarop dui dat ten minste een woord tydens die skep van 'n samesmelting in sy volledige vorm is. Betreffende die nuwe woorde met streeketikette is byna 80% geëtiketteer as behorende tot Britse en Noord-Amerikaanse gebruike, 'n aanduiding van die verskuilde Anglo-sentrisiteit. Die status van sommige woorde wat as neologismes beskou word, is betwisbaar, aangesien hulle lank reeds gebruik word. Die impak van die wetenskap en tegnologie op die insluiting van neologismes in OALD 10 is taamlik duidelik. Daar is steeds heelwat ruimte vir verdere navorsing rakende OALD 10.

Sleutelwoorde: NEOLOGISME, OALD 10, WOORDSOORT, WOORDVORMING, STREEKETIKET, ANGLO-SENTRISITEIT, WOORDEBOEKHERSIENING, WETENSKAPLIKE EN TEGNIESE INVLOED

1. Introduction

Neologisms almost spring up on a daily basis. So do new senses of the current words. As the record of a language, dictionaries need to reflect such a trend as timely as possible. In doing so, neologisms and new senses will be included in dictionaries, while those obsolete or outdated words and senses will be deleted. This is also true of learner's dictionaries. English learner's dictionaries, including the "Big Five", have had a history of timely revision. This is due to the fact that their compilers are well aware that dictionary contents start to become outdated the moment they come to print (Cheng and Liu 2019: 62). Timely revision can help make a dictionary up to date, thus helping attract the potential buyers and boost the sales of dictionaries.

The revision of the 10th edition of the printed *Oxford Advanced Learner's Dictionary* (henceforth OALD 10) can be a case in point. In recent years, the frequency of its revision spans about 5 years. For example, OALD 9 was published in 2015 while OALD 10 in 2020. The compilers of OALD 10 added 1204 neologisms, among other things, in revising OALD 9. Technically speaking, a neologism is "a new word or expression, or a new meaning of a word" (Lea and Bradbery 2020: 1044). However, in this article, it only refers to the words or expressions newly added to the headword list of OALD 10. The inclusion of neologisms will be discussed in terms of their part-of-speech labeling, word-building, regional labeling, and so on. Doing so can also shed light on how to revise a dictionary of similar or other types in terms of integrating neologisms. Neologisms or new

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1871 (Lexiconote)

words will be used interchangeably below to refer to those newly added words or expressions in OALD 10 hereafter.

2. On the inclusion of neologisms in OALD 10

In the following subsections, the inclusion of neologisms will be discussed from the perspectives of part-of-speech labeling, word-formation, and regional labeling respectively.

2.1 Part-of-speech analysis of neologisms

All the 1204 new words included have been classified in terms of their part of speech as presented in Table 1. They are arranged, percentagewise, in the decreasing order. It needs to be explained here that "suffix", "idiom", "combining form" and "abbreviation" cannot be labeled with parts of speech. However, to present a holistic picture of part-of-speech labeling of the neologisms included in OALD 10 to readers, they are still included in Table 1.

Table 1: Part of speech analysis of neologisms

Part of speech	Total	Percentage
noun	891	71.1%
adjective	175	14%
adverb	46	3.7%
abbreviation	42	3.4%
verb	20	1.6%
noun & adjective	10	0.8%
exclamation	7	0.6%
noun & verb	6	0.5%
adjective & adverb	2	0.2%
preposition	1	0.08%
interjection	1	0.08%
suffix	1	0.08%
idioms	1	0.08%
combining form	1	0.08%

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1871 (Lexiconote)

The fact that nouns figure predominantly among 1204 neologisms integrated into OALD 10 could help speak for the preference of English of using nouns (cf. Lian 1993: 105, and Liu 2010: 3, on the nouny nature of English). According to Table 1, nouns add up to 891, and account for over 71.1% of the total neologisms. Heterosemy, with a word serving more than one part of speech, is common in many languages, including English. Among the neologisms, 10 words are both nouns and adjectives, while 6 are nouns and verbs simultaneously. If the 16 nouns are added to 891, the percentage of nouns will be further increased. The strikingly low percentage of verbs among the neologisms could also help support such nature. In other words, the large percentage of nouns among the neologisms helps demonstrate the static rather than the dynamic nature of English.

Several other simple facts regarding the part-of-speech labeling of the neologisms need to be noted as well. The percentages of adjective, adverbs and verbs are larger than 1%, coming in the decreasing order. Nouns, adjectives, adverbs and verbs account for almost 96% of the neologisms included. In other words, the over-majority of neologisms are content words rather than functional ones. As for exclamations, interjections and prepositions, judging by their percentage, they are so few, and are almost negligible. In the meantime, in addition to 42 abbreviations, there are also other lexical items included therein, including one "suffix" (-es), one "idiom" (betcha), and one "combining form" (-oriented). However, all those items can hardly qualify as a "word" in its traditional sense, even if they are included in the headword list in OALD 10. It is almost impossible to analyze them in terms of part-of-speech. However, the relatively high percentage of abbreviations proves that since the beginning of this century, abbreviations have increased greatly with the wide use of internet (Gao 2022).

2.2 Classifications of neologisms according to word-formation

All the neologisms included in OALD 10, except "-es" and "-oriented", can be analyzed in terms of their word-formation. Generally speaking, the main methods for creating a new word are compounding and derivation. Shifting or conversion, one of the three major word-formations, is an important way to generate a new sense of a neologism (Gao 2020: 52). In the following section, discussions will center on the neologisms formed through compounding, derivation and blending, since the neologisms formed in the three ways make up nearly 80% of those included in OALD 10. As for other ones formed in other ways, they are simply mentioned in passing.

2.2.1 Compounding

Compounds constitute more than half of the neologisms included in OALD 10, which can appear as open, tight or hyphenated ones. 623 neologisms are compounds, accounting for 51.7% of the total. This provides strong evidence for

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1871 (Lexiconote)

compounding to be "the most frequently used way of creating neologisms" (Gao 2021: 25). According to OALD 10, a compound refers to "a noun, an adjective or a verb made of two or more words [...] written as one or more words, or joined by a hyphen. Travel agent, dark-haired and bathroom are all compounds" (Lea and Bradbery 2020: 311). Comparatively speaking, a compound creates less cognitive load for a language user than a newly coined word, which helps to explain why compounds figure so prominently among the neologisms in OALD 10.

Compounds, as suggested by Lea and Bradbery (2020), can consist of multiword forms, hyphenated words or two or more words combined to orthographically form one word. Words like "angel investor", "attitude problem", "family values", can serve as multiword forms. These compounds add up to 513, accounting for 82.3% of the total compounds in OALD 10. Hyphenated compounds, like "cash-rich" and "nail-biter" total to 62, which accounts for 10% in OALD 10. Compounds consisting of two or more words spelt as one, such as "wingsuit" and "upvote", total 48 altogether, which takes up 7.7% in OALD 10. Just like nouns figure prominently in the neologisms included, the over-majority of compounds are nouns. The total adjective and adverbial compounds are so few, thus negligible.

2.2.2 Derivation

Derivation is one of the major ways by which the neologisms in OALD 10 are coined. It is also a major way of creating a new word, which attaches an affix to an existing word. According to Table 2, it can be seen that 300 words are coined by derivation, accounting for almost 24.9% of all new words incorporated into OALD 10. However, it needs to be pointed out that 170 neologisms are already listed as derivatives in OALD 9, accounting for 56.7% of all derivatives. They are counted as neologisms for the reason that they are listed among the headwords. Meanwhile, 130 neologisms are not, which accounts for 43.3% of all derivatives. Among the derivatives, the suffixes used most are "-ly", "-ed" and "-ing". The former is mostly used to make an adverb, while the latter two are for making adjectives. The top three prefixes for coining the neologisms are "un-", "self-" and "non-".

Table 2: Analysis of neologisms using derivation

Neologisms appearing as derivatives in OALD 9	Neologisms not appearing as derivatives in OALD 9
170	130
56.7%	43.3%

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1871 (Lexiconote)

2.2.3 Blending

Apart from compounding and derivation, the third major way to make the neologisms in OALD 10 is blending. In total, 34 new words are coined by blending, summarized in Table 3. In Table 3, beginning and end refer to the beginning and the end of a word respectively.

Table 3: Analysis of the neologisms formed by blending

How a blend is formed	Total	Percentage
Word+end	5	14.7%
Word+beginning	5	14.7%
Beginning+word	16	47.1%
Beginning+beginning	7	20.6%
Beginning+end	1	2.9%

According to Table 3, five ways have been adopted to coin a blend in OALD 10. Nearly half of the blends are coined with the beginning part of a word mixing with a word, such as "Brexit", "e-publishing" and "alter-right". 5 blends are made up of a word mixing with the ending part of another one, including "mansplain", "hackathon", and so on, while another 5 a word mixing with the beginning part of another one, such as "Big Pharma" and "sales rep". In the meantime, 7 blends are made up of the beginnings of two words, like "prelim". Only one blend, "vlog", is made up of the beginning of "video" and the end of "blog". It can be seen clearly that over 75% of the blends above are partial in nature. That is, at least one word to make a blend is in its full form.

2.3 Classification of neologisms in terms of regional labels

Neologisms with regional labels in OALD 10, in terms of their distribution, clearly displays the trace of Anglo-centrism. That is, the majority of the new words with regional labels belong to British and North American usages, while those from other English-speaking regions make up less than 20% of the total with such labels. Altogether, there are 278 words with regional labels in OALD 10, accounting for 24.6% of the total neologisms. Table 4 shows the distribution of those neologisms with the regional labels.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1871 (Lexiconote)

Table 4: The distribution of neologisms with regional labels

Regional label	Total	Percentage
***NAmeE* (including 16 US and 2 Canadian usages)**	168	60.4%
BrE	50	18%
AustralE	4	1.4%
***AfrE*（including 11, 18 and 7 East, South and West African English usages）**	36	12.9%
***AsianE* (including 16 Indian and 4 East-Asian English usages)**	20	7.3%

According to Table 4, 150 words belong to North American English (*NAmeE*), making up 54% of those with regional labels. However, there are another 16 words labeled US usages and 2 Canadian ones. Therefore, the total neologisms belonging to North American usages add up to 168, accounting for 60.4% of those with regional labels. Besides, another 50 neologisms are labeled British English (*BrE*), accounting for 18%. There are another 4 from Australian English. 11, 18 and 7 neologisms are labeled as belonging to East, South and West African English respectively. Those labeled with *AfrE* constitute 12.9% of the total with regional labels. 16 Indian English neologisms account for 5.8% of those with regional labels, while 4 South-East Asian ones account for 1.5% of the total. It can be seen that among the neologisms with regional labels, the over-majority of them are from countries with English as the first language, predominantly British and North American English.

It is true that dictionary compilers should stay as neutral as possible and should not show personal bias in compiling a dictionary. This is also true of the compilers of learner's dictionaries like OALD 10. Even if the compilers of OALD 10 may claim to be unbiased in the inclusion of new words with regional labels, a closer look at those words has revealed a different picture to us. This is quite similar to the fact that Dr. Johnson's definition of oats as "A grain, which in England is generally given to horses, but in Scotland supports the people" (Johnson 1755, cited in Tian 2017), which reveals him to be quite biased. Therefore, it can be argued that in the selection of the neologisms to be included in OALD 10, the socio-economic impact of a country or region may have played a very important role. The distribution of the neologisms with regional labels in Table 4 could reveal the hidden Anglo-centrism in integrating neologisms into OALD 10. This could have something to do with the socio-economic development and political power of those countries with English as their mother tongue. However, it can still be argued that the inclusion of Asian and African English neologisms can be further increased so as to reveal English to be true "world English".

3. Two more points worthy of being noted

3.1 Should those words be counted as neologisms?

As has been pointed out in the beginning of the article, quite a number of words or expressions newly integrated into the headword list of OALD 10 are not neologisms in their true sense. A closer look reveals that actually a lot of the words or expressions have been in use for a very long time and are not new at all. They make the study objects of the present study simply because they are integrated into the headword list for the first time in OALD 10. For example, "*Achoo*", an interjection, according to *Longman Dictionary of Contemporary English* (https://www.ldoceonline.com/dictionary/achoo), has been coined between 1800 and 1900. This also applies to "*-es*", plural or third person singular marker, which has long been in use. Combining form "*-oriented*", meaning "directed towards something or made or adapted for a particular purpose" (Lea and Bradbery 2020), may have been used by people for a long time as well, since there are more than 20 compounds relevant to it.

Therefore, when dictionary compilers claim they have integrated a certain number of new words and senses into a dictionary in revising it, the potential buyers may need to take a second thought. The reason is that the compilers and the buyers stand on different grounds. For the compilers, they may stand with publishers. Therefore, the exaggerated neologisms integrated could make the biggest selling point to attract the potential buyers. However, for dictionary buyers, one of the main reasons to draw them to buying the revised edition of a dictionary could be the neologisms in their true sense. For them, it is such neologisms that keep the dictionary up to date.

Meanwhile, some words have not been counted as new words, since they are just the minor modifications of those already present in the headword list in OALD 9. Such words add up to 50. In this study, they are excluded from analysis. The modifications mainly include the deletion or addition of "the", "hyphen", plural marker like "s" or "es", etc. For example, in OALD 10, "*Top ten*" merely deletes "the", while "the" is added to "Med", the clipping of the Mediterranean Sea, thus making "*The Med*". "*rough-and-ready*" removes the hyphens to be "*rough and ready*", but a hyphen is added to "*roommate*" to make "*room-mate*". The plural form "*school days*" in OALD 9 is changed into "school day" in OALD 10. And "*luvvy" in* OALD 9 *in changed into* "*luvvie*" in OALD 10. Words like them can hardly justify themselves as neologisms newly integrated into OALD 10, thus excluded from analysis.

3.2 The influence of the sci-tech on the coinage of neologisms

It is understandable that with the development of human society and the advances in science and technology, esp. internet technology, a lot of words are coined, including "EdTech" (education technology), "e-business" (electronic business),

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1871 (Lexiconote)

"vlog" (video blog) and so on. This is well reflected in the coverage of neologisms in OALD 10.

One of the features that people are influenced by technology, especially Internet technology, is that abbreviations or informal words or phrases are used, thus introduced into OALD 10 accordingly. For example, *thanx* (thanks), *plz* (please), and *BFF* (best friend forever) are all informal form of their original ones and are often used in email or online social platforms. One of the possible reasons that people use such writing is that they want to communicate casually with their online pals. And this writing would be fast by using abbreviations and even emoji(s) to express their feeling. Anyway, such neologisms find their way into the list of headwords.

There are also other neologisms which are more evidently relevant to technological advances, "live stream" and "live feed" can serve as cases in point. The former means "a live broadcast of an event over the internet" (Lea and Bradbery 2020: 920), while the latter means "the broadcast of sound or video over the internet from a live (not recorded) source, for example a concert or sports event" (Lea and Bradbery 2020: 920). The inclusion of such words in the headword list implies the potential impact of technology on people, especially internet culture.

The impact of science and technology on neologisms can also be reflected in the conversion of part of speech in newly added words. Google and WhatsApp™ can help illustrate the view very well. People begin to add the verb usage to the name of IT companies or apps. For example, Google, a well-known search engine for netizens, derives its verb usage of using Google to search for someone or something. And WhatsApp™, an app that allows people to chat through texting messages, sending photos or short videos, also develops its own verb usage, i.e., to use WhatsApp to chat with someone.

4. Conclusion

In this article, the neologisms included in OALD 10 have been approached from different perspectives. In terms of the part of speech, the over-majority are nouns, which could speak for the nouny nature of English. Judging by the word-formation, the top three ways to form the neologisms are compounding, derivation and blending. Compounds consisting of more than one word seem to take precedent over the hyphenated or one-word forms. Most derivatives are also nouns. However, more than half of the neologisms formed by derivation have appeared as derivatives in OALD 9. For blends, most of them are partial in nature, which means they include one full word. When it comes to the new words with regional labels, over 80% are labeled as belonging to British and North American usages, a sign for the hidden Anglo-centrism in integrating such words. Actually, the status of some words as "new words" is questionable since they have been in use for quite some time. In addition, the impact of science and technology on the inclusion of neologisms in OALD 10 is quite great.

The present study can be expanded in different ways. One can conduct a comparative study of the inclusion of neologisms in OALD 10 and other English learner's dictionaries to see what similarities or differences may surface in terms of their principles for guiding the inclusion of the neologisms. In addition, most neologisms are furnished with examples. It would be interesting to investigate to see if the principles for providing examples for the neologisms are the same as those for the existing words in the dictionary. Besides, apart from neologisms, various other aspects regarding the revision of OALD 10 also deserve our attention, such as definition, examples, outside matter.

Acknowledgement

The Research Program is supported by Science Foundation of Guangxi Minzu University (2021SKQD32)

References

Cheng, Zhaowei and Hua Liu. 2019. A New Concise English–Chinese Dictionary of Science and Technology: Revision and Innovation. *Chinese Science & Technology Translators Journal* 32(2): 62-66.

Gao, Yongwei. 2020. Discussion of Neologisms Related to Cryptocurrency and Their Translation. *China Terminology* 22(6): 51-56.

Gao, Yongwei. 2021. *A Study on Word-Formation in Contemporary English*. Shanghai: Shanghai Translation Publishing House.

Gao, Yongwei. 2022. *A Dictionary of Acronyms and Initialisms in Contemporary English*. Shanghai: Shanghai Foreign Language Education Press.

Hornby, A.S. 2015. *Oxford Advanced Learner's Dictionary* 9th Edition. Oxford: Oxford University Press.

Lea, Diana and Jennifer Bradbery. 2020. *Oxford Advanced Learner's Dictionary* 10th Edition. Oxford: Oxford University Press.

Lian, Shuneng. 1993. *Contrastive Studies of English and Chinese*. Beijing: Higher Education Press.

Liu, Danqing. 2010. Chinese as a Verby Language: On Typical Differences between Verby Languages and Nouny Languages. *Chinese Teaching in the World* 1: 3-17.

Tian, Bing. 2017. A Case Study of Philological and Special Style in Defining the Plant Nouns in Samuel Johnson's *Dictionary of the English Language* (1755). *Foreign Language and Literature* 1: 104-109.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1879 (Article)

Der Effizienz- und Intelligenzbegriff in der Lexikographie und künstlichen Intelligenz: kann ChatGPT die lexikographische Textsorte nachbilden?

Iván Arias-Arias, *Instituto da Lingua Galega, Universidade de Santiago de Compostela, Spanien (ivanarias.arias@usc.gal) (https://orcid.org/0000-0003-2673-0899)*

María José Domínguez Vázquez, *Instituto da Lingua Galega, Universidade de Santiago de Compostela, Spanien (majo.dominguez@usc.es) (https://orcid.org/0000-0002-6060-9577)*
und
Carlos Valcárcel Riveiro, *Universidade de Vigo, Spanien (carlos.valcarcel@uvigo.gal) (https://orcid.org/0000-0003-1123-5211)*

Zusammenfassung: Mittels Pilotexperimente für das Sprachenpaar Deutsch–Galicisch untersucht der vorliegende Aufsatz den Effizienz- und Intelligenzbegriff in der Lexikographie und künstlichen Intelligenz (KI). Die Experimente versuchen, empirisch und statistisch fundierte Erkenntnisse über die lexikographische Textsorte „Wörterbuchartikel" in den Antworten von ChatGPT-3.5 zu gewinnen, und darüber hinaus über die lexikographischen Daten, mit denen dieser Chatbot trainiert wurde. Zu diesem Zweck werden sowohl quantitative als auch qualitative Methoden herangezogen. Der Analyse liegt die Auswertung der Outputs von mehreren Sessions mit demselben Prompt in ChatGPT-3.5 zugrunde. Zum einen wird die algorithmische Leistung von intelligenten Systemen im Vergleich zu Daten aus lexikographischen Werken bewertet; zum anderen werden die gelieferten ChatGPT-Daten über konkrete Textteile der genannten lexikographischen Textsorte analysiert. Die Resultate dieser Studie tragen dazu bei, nicht nur den Effizienzgrad von diesem Chatbot hinsichtlich der Erstellung von Wörterbuchartikeln zu evaluieren, sondern auch in den Intelligenzbegriff, die Denkprozesse und die in beiden Disziplinen auszuführenden Handlungen zu vertiefen.

Stichwörter: LEXIKOGRAPHIE, KI, CHATGPT-3.5, WÖRTERBUCHARTIKEL, EFFIZIENZBEGRIFF, INTELLIGENZBEGRIFF, LEXIKOGRAPHISCHE TEXTSORTE, TRAININGSDATEN, LEXIKOGRAPHISCHE DATEN

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1879 (Article)

Abstract: Efficiency and Intelligence in Lexicography and Artificial Intelligence: Can ChatGPT Recreate the Lexicographical Text Type? By means of pilot experiments for the language pair German–Galician, this paper examines the concept of efficiency and intelligence in lexicography and artificial intelligence (AI). The aim of the experiments is to gain empirically and statistically based insights into the lexicographical text type "dictionary article" in the responses of ChatGPT-3.5, as well as into the lexicographical data on which this chatbot was trained. Both quantitative and qualitative methods are used for this purpose. The analysis is based on the evaluation of the outputs of several sessions with the same prompt in ChatGPT-3.5. On the one hand, the algorithmic performance of intelligent systems is evaluated in comparison with data from lexicographical works; on the other hand, the ChatGPT data supplied is analysed using specific text passages of the aforementioned lexicographical text type. The results of this study not only help to evaluate the efficiency of this chatbot regarding the creation of dictionary articles, but also to delve deeper into the concept of intelligence, the thought processes and the actions to be carried out in both disciplines.

Keywords: LEXICOGRAPHY, AI, CHATGPT-3.5, DICTIONARY ARTICLE, CONCEPT OF EFFICIENCY, CONCEPT OF INTELLIGENCE, LEXICOGRAPHICAL TEXT TYPE, TRAINING DATA, LEXICOGRAPHICAL DATA

1. Einführung

In den letzten Jahrzehnten hat sich die Künstliche Intelligenz (KI) explosionsartig entwickelt, so dass sogar einige mit ihr verbundenen Entwicklungen Bestandteil unseres alltäglichen Lebens geworden sind. Somit kennen oder verwenden wir alle z.B. digitale Assistenten wie *Alexa*, *Siri* oder *Google Assistant*. Der Einfluss der KI auf unseren Alltag ist unbestritten, auch wenn wir uns dessen häufig nicht bewusst sind; ihr Einfluss auf die lexikographische Tätigkeit hat eine rege Diskussion ausgelöst. Manche Autoren sind der Ansicht, dass der Output von KI-Tools wie dem Chatbot ChatGPT-3.5[1] (*Generative Pretrained Transformer*, https://chat.openai.com/) statistisch gesehen die Ergebnisse einiger Wörterbücher übertrifft (vgl. Phoodai und Rikk 2023) und manche prophezeien sogar das Ende der Lexikographie (vgl. de Schryver und Joffe 2023).

In diesem Zusammenhang geht es hier um die Frage der Effizienz (s. 3) von ChatGPT bei einem konkreten Prompt[2] (Eingabeaufforderung) betreffend den Wörterbuchartikel als lexikographische Textsorte (vgl. Wiegand 1996). Konkret fragen wir, inwiefern die ChatGPT-Ergebnisse für das Deutsche und das Galicische mit der Mikrostruktur von Referenzwörterbüchern (DUDEN-Onlinewörterbuch, https://www.duden.de/woerterbuch, fortan DUDEN und Dicionario da Real Academia Galega, https://academia.gal/dicionario/rag, fortan DRAG) quantitativ bzw. qualitativ vergleichbar sind. Eine Analyse der lexikographischen Daten sowie ihr Vergleich mit denen in Referenzwörterbüchern ist unseres Erachtens die erste Frage, die überhaupt gestellt werden sollte. Daher stehen die Datenverfügbarkeit sowie ihre Gegenüberstellung im

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1879 (Article)

Mittelpunkt, jedoch nicht die Analyse der möglichen Verwendung von ChatGPT bei lexikographischen Aufgaben.

Dazu gliedert sich der Aufsatz wie folgt: in Kapitel 2 wird ein Gesamtüberblick über ausgewählte Schwerpunkte der Lexikographie und der KI dargestellt. Dem Effizienz- und der Intelligenzbegriff widmet sich einführend Kapitel 3: die Frage lautet, ob ChatGPT über Daten bezüglich der Textsorte „Wörterbuchartikel" verfügt. Kapitel 4 dient der Erklärung der Methode und bietet die statistisch ausgewertete Untersuchung der Ergebnisse aus dem Chatbot an. Die Hauptergebnisse der quantitativ und qualitativ ausgerichteten Analyse werden in Kapitel 5 vorgestellt. Schlussfolgerungen werden in Kapitel 6 gezogen.

2. Zur Lexikographie und zur künstlichen Intelligenz: Einführendes

Das WLWF-3 (2020: 224) definiert Lexikographie als „Menge aller Aktivitäten, die auf die Erstellung lexikographischer Nachschlagewerke gerichtet ist". Neben der theoretischen Untersuchung auf Gebieten wie der Metalexikographie, Wörterbuchforschung, Wörterbuchkritik u.a. und dem Erwägen und der Anwendung verschiedenartiger Strategien, Methoden und Techniken zur Analyse des Sprachmaterials befasst sich die lexikographische Tätigkeit mit der Entwicklung von Informationssystemen (vgl. Villa Vigoni 2018, https://www.emlex.phil.fau.de/ueberuns/publikationen/andere-publikationen/), in jeglichem Format und für verschiedenartige Zugangsgeräte (Wörterbücher, Portale, Wörterbuch-Apps u.a.), bei denen i.d.R. menschliche Benutzende Antwort auf eine Suchanfrage finden können (s. 5). Sowohl die adäquate Datenpräsentation angesichts der Wörterbuchziele und -adressaten, als auch die Datenqualität bzw. Verantwortung für ihre Qualität (Kouassi 2022) stehen als lexikographische Aufgaben im Mittelpunkt. Die Lexikographie wird zudem als eine wissenschaftliche und kulturelle Praxis aufgefasst, die die Entstehung der Werke sowie ihren Gebrauch ermöglichen sollte (Wiegand 1983: 38), und dabei trägt sie auch eine gesellschaftliche Verantwortung. Die Rolle des Wörterbuchs als Autorität ist in diesem Zusammenhang nicht zu übersehen (Kosem et al. 2019).

Hinsichtlich des anvisierten Benutzerkreises lassen sich die Endprodukte der Lexikographie in Bezug auf Sprachniveaus, Ziele u.a. klassifizieren, aber auch bezüglich der Unterscheidung zwischen Computerlexikographie und computergestützter Lexikographie. In diesem Zusammenhang kann man behaupten, dass Wörterbücher für menschliche BenutzerInnen die primären Endprodukte der computergestützten Lexikographie sind (s. 5). Hingegen befasst sich die Computerlexikographie mit der Entwicklung von Ressourcen für die maschinelle Weiterverarbeitung. Solche maschinenlesbaren Lexika[3] finden Anwendung in verschiedenen NLP-Aufgaben wie *Parsing*, maschineller Übersetzung oder Sprachgenerierung. Folglich stellen sich Maschinen als primäre

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1879 (Article)

Adressaten (nicht BenutzerInnen) der Ergebnisse der Computerlexikographie heraus (Weiteres dazu in 5).

Daraus folgt, dass die Zusammenstellung von lexikographisch akkuraten Daten zwecks ihrer maschinellen Lesbarkeit auch als eine mögliche Aufgabe der Lexikographie hervortritt. Eigentlich ist das nicht neu, denn schon Mel'čuk (1984) hat sein *Dictionnaire explicatif et combinatoire du français contemporain* mit Blick auf diese mögliche Anwendung der lexikographischen Daten entwickelt.

Im Gegensatz zur Lexikographie ist die KI ein Teilgebiet der Informatik, das sich mit der maschinellen Nachahmung menschlicher Intelligenz befasst, d.h. mit der Entwicklung intelligenter Systeme (Mainzer 2019), die bei der Durchführung unterschiedlicher Aufgaben Aspekte des menschlichen Verhaltens wie der Problemlösung oder der Entscheidungsfindung simulieren (McCarthy 2007). Die KI hat viele Anwendungsbereiche, dementsprechend spielt sie eine wesentliche Rolle bei vielen Prozessen und den dadurch geschaffenen Endprodukten, wie z.B. bei der Entwicklung autonomer Fahrzeuge oder Roboter (z.B. für die Erforschung des Weltraums), bei den medizinischen Diagnosen, bei der Marktforschung für ein Produkt, bei der Bilderkennung (Gesichtserkennung oder Objekterkennung), u.a. Die natürliche Sprachverarbeitung (NLP), die zur Datenverarbeitung und -interpretation der menschlichen Kommunikationsprozesse Computeralgorithmen und maschinelles Lernen heranzieht, gilt als Teilgebiet der KI. Als besonders relevant erweisen sich die Sprachmodelle, da sie in unterschiedlichen Bereichen eingesetzt werden können bzw. als Grundlage unterschiedlicher Produkte und Ressourcen dienen können. Im Weiteren werden einige genannt:

- Sprachassistenten wie *Alexa, Siri* oder *Google home*.
- Chatbots und virtueller Kundenservice: Es handelt sich um Softwareanwendungen bzw. virtuelle Assistenten, die häufig gestellte Fragen beantworten. Man strebt z.B. dabei an, den Kundenservice zu verbessern (Banken, Online-Shops, u.a.). Sie werden auch für den Querverkauf von Produkten, als Nachrichten-Bots beim elektronischen Handel oder als Nachrichten-Apps wie *Facebook Messenger*. Als Chatbots lassen sich ChatGPT (s. 3 und 4), *Bing CHAT* (https://www.bing.com/) oder *Bard* (https://bard.google.com/) nennen. Mit diesen Dienstleistungen und Werkzeugen stehen Untersuchungen zur Analyse der Meinungsforschung oder der Analyse von Gefühlen in engem Zusammenhang: ihre Ergebnisse finden eine direkte Anwendung in Gebieten wie dem elektronischen Handel oder der Meinungsanalyse in den sozialen Netzwerken, indem sich ein Text oder ein nutzergenerierter Inhalt durch eine automatische maschinelle Analyse als positiv, negativ oder neutral bewerten lässt. Einige Beispiele dazu sind *Linguakit* (https://linguakit.com/es/analizador-de-sentimient) oder *SentiWordNet* (https://github.com/aesuli/SentiWordNet).

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1879 (Article)

- Roboter für die schriftliche Textproduktion, wie *z.B. Inferkit* (https://inferkit.com/), *Sassbook AI* (https://sassbook.com/), oder AI-Schreibsysteme, wie *Rytr* (https://app.rytr.me). Insgesamt erstellen sie automatisch Texte aus einer reduzierten Anzahl an Wörtern als Ausgangspunkt.
- Weitere Generatoren sind vorhanden:
  - Für die Erstellung von sportlichen Berichten, Zusammenfassungen, vereinfachten Texten usw. liegen Beispiele vor (Nallapati et al. 2016 oder Roemmele 2016).
  - Für die Wiedergabe mündlicher Sprache und Interaktionen im Gespräch, wie z.B. Systeme, die mit der Stimme einer Person trainiert werden und Text produzieren bzw. Gespräche führen. Es gibt Apps wie *Pi.ai/talk* (https://pi.ai/talk), *Call Annie* (https://callannie.ai) oder *character.ai* (https://beta.character.ai/; für Jugendliche), eine Art KI-Freunde bzw. intelligente Assistenten.
  - Für die Erstellung von Bildern nach Beschreibungen in natürlichen Sprachen, wie *DALL-E* (https://openai.com/dall-e-2), von Open AI.
  - Automatische Übersetzung, wie z.B. *WIPO Pearl* (https://wipopearl.wipo.int/en/linguistic).

Unausgesprochen weisen die KI und die Lexikographie andere Endprodukte und Ziele auf. Eine weitere Entscheidung sei hier hervorzuheben: Sprachmodelle streben an, dass Maschinen nachahmen, wie Menschen Wörter verwenden, d.h. die menschliche Kommunikation; die KI hingegen hat die Entwicklung von Maschinen und KI-Tools als Ziel, die intelligent agieren. Die Wechselwirkungen sind nicht zu übersehen.

3. Effizienz und Intelligenz am Beispiel von ChatGPT

An erster Stelle scheint es sinnvoll, zu erläutern, warum sich der vorliegende Beitrag mit ChatGPT befasst und warum wir ChatGPT als intelligentes System auswählen. Benutzungsstudien (vgl. Domínguez Vázquez und Valcárcel Riveiro 2015 oder Müller-Spitzer und Koplenig 2015) stellen fest, dass BenutzerInnen Faktoren wie (a) eine schnelle und bequeme Abfrage sowie (b) einen leichten und kostenfreien Zugang schätzen. ChatGPT erfüllt diese Voraussetzungen, und deswegen kann vorausgesagt werden, dass BenutzerInnen es einem Wörterbuch bevorzugen würden. Dies lässt sich auch dadurch begründen, dass ChatGPT in nur fünf Tagen die Grenze von einer Million NutzerInnen überschritten hat (s. Abbildung 1).

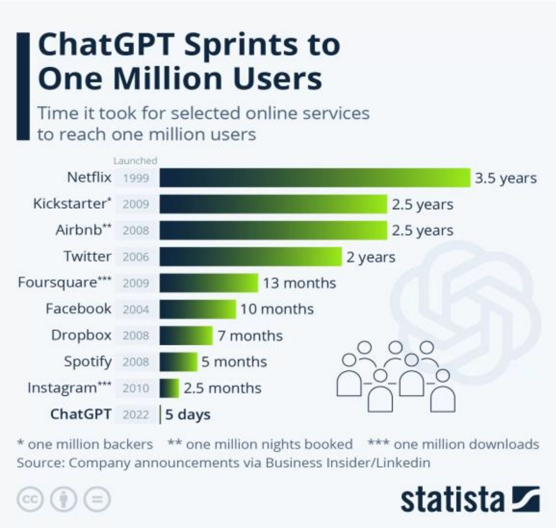


Abbildung 1: Zeit, die Online-Dienste bis zum Erreichen von 1 Million NutzerInnen benötigten

ChatGPT ist ein Chatbot, der aus einem Sprachmodell mit generativen Vortraining stammt und ein autoregressives System enthält, d.h., dass es Tokens vorhersagt, sie zum Prompt hinzufügt und sie wieder in das Modell einspeist. Das ist der Grund, weshalb sich solche Modelle als sehr effizient erwiesen haben, wenn die Eingabeaufforderung darin besteht, Texte oder Textsorten zu (re)produzieren (vgl. Radford et al. 2018). Die Grundannahme der hier zu präsentierenden Pilotexperimente basiert darauf, dass ChatGPT — wie andere generative Sprachmodelle — menschenähnliche Texte von hoher Qualität generieren kann. Ob das in der Tat bei der Erstellung von Wörterbuchartikeln im Deutschen und im Galicischen so ist, ist Gegenstand dieser Untersuchung (s. 4).

In diesem Zusammenhang sei hier auf den Intelligenzbegriff hinzuweisen, der laut der Wikipedia-Definition mit der Anwendung von kognitiven Fähigkeiten zusammenhängt:

> Intelligenz (von lateinisch *intellegere* „erkennen“, „einsehen“; „verstehen“; wörtlich „wählen zwischen …“ von lateinisch *inter* „zwischen“ und *legere* „lesen, wählen“) ist die kognitive bzw. geistige Leistungsfähigkeit speziell im Problemlösen. Der Begriff umfasst die Gesamtheit unterschiedlich ausgeprägter kognitiver Fähigkeiten zur Lösung eines logischen, sprachlichen, mathematischen oder sinnorientierten Problems.

Hier kommt die Frage auf, was das Wort „Intelligenz“ im Konzept „künstliche Intelligenz“ zu bedeuten hat. Dazu äußert sich McCarthy (2007: 2) wie folgt:

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1879 (Article)

> It is the science and engineering of making intelligent machines, especially intelligent computer programs. It is related to the similar task of using computers to understand human intelligence, but AI does not have to confine itself to methods that are biologically observable.

McCarthy (2007: 2) beschreibt diese Intelligenz als die Fähigkeit, bestimmte Ziele zu erreichen. Diese Beschreibung bringt eine weitere Frage mit sich, und zwar, was intelligente Maschinen sind oder inwiefern sie als solche aufgefasst werden können. Darauf antwortet McCarthy (2007: 3), dass die Maschinen nicht intelligent sind, aber er ergänzt seinen Ansatz wie folgt:

> Intelligence involves mechanisms, and AI research has discovered how to make computers carry out some of them and not others. If doing a task requires only mechanisms that are well understood today, computer programs can give very impressive performances on these tasks. Such programs should be considered "somewhat intelligent".

Da der Intelligenzbegriff mit dem Verfahren zur Problemlösung in engem Zusammenhang steht, ist aus computergestützter Perspektive auf die Verbindung zwischen Intelligenz und System näher einzugehen. Folglich definiert Mainzer (2019: 3) intelligente Systeme folgenderweise:

> Ein System heißt intelligent, wenn es selbständig und effizient Probleme lösen kann. Der Grad der Intelligenz hängt vom Grad der Selbstständigkeit, dem Grad der Komplexität des Problems und dem Grad der Effizienz des Problemlösungsverfahrens ab.

Auf die drei genannten Intelligenzgrade eines Systems wird jetzt eingegangen:

(a) Der Grad der Selbständigkeit hängt von dem zu benutzenden System ab. Eine Computeranwendung kann selbständig sein, wenn sie unabhängig von anderen Ressourcen Probleme lösen kann.

(b) Der Grad der Komplexität des Problems hängt von der Handlung selbst ab. In diesem Sinne ist mit einer bunten Vielfalt von Situationen zu rechnen, die den Komplexitätsgrad erhöhen können. Hier geht man davon aus, dass die Benutzenden ein System zu Rate ziehen sollen, das zur Beantwortung ihrer Anfragen geeignet ist; z.B. ist von einem Korpus nicht zu erwarten, dass es Definitionen liefert, wie von einem Wörterbuch nicht verlangt werden kann, dass es automatisch Texte generiert.

(c) Der Grad der Effizienz des Lösungsverfahrens hängt davon ab, ob ein System in der Tat ein konkretes Problem lösen bzw. eine konkrete Frage angemessen beantworten kann. Bei der Bestimmung der Effizienz sind der Wirkungsgrad (s. 5) und die Rolle der BenutzerInnen (s. 5) von entscheidender Bedeutung.

Seitens der Lexikographie wird der Effizienzbegriff klar definiert. Laut Wiegand (1998: 259) ist die Benutzungseffizienz als Resultat einer Adäquatheit des Produk-tes (des Wörterbuches) an die Bedürfnisse der BenutzerInnen zu verstehen, wozu die wissenschaftlichen Kenntnisse der Benutzenden und die Etablierung der Konsultation eines Referenzwerkes als kulturelle Praxis einen großen Beitrag leisten. Im Grunde genommen haben beide — die Lexikographie und die KI — ein gemeinsames Prinzip: Man ist auf der Suche nach Lösungen für ein Problem und benötigt dazu Mechanismen und Strategien. Im Gegensatz zu Menschen stützen sich Maschinen jedoch auf Theorien menschlicher Denk- und Lernprozesse, und insbesondere auf die Nachbildung dieser Prozesse (vgl. 5). In diesem Zusammenhang sei daran zu erinnern, dass die KI eine unvorstellbare Menge an Daten analysieren kann und in der Lage ist, Muster und Daten auf der Grundlage menschlicher Kommandos zu reproduzieren. Die konkrete Frage lautet dann, inwiefern ein Gespräch mit einem intelligenten System effizient zur Lösung eines sprachlichen bzw. lexikographischen Problems führen kann, und das hängt vom Effizienzgrad des Lösungsverfahrens ab. Den Effizienzbegriff verstehen wir daher in der vorliegenden Arbeit auf zwei Weisen: zum einen spricht die Datenmenge, über die ein System oder ein Nachschlagewerk verfügen, gewissermaßen für einen höheren Wirkungsgrad (s. 5.4); zum anderen gilt der Benutzertyp als Voraussetzung für die Effizienz des Problemlösungsverfahrens (s. 5.1 und 5.4). Begriffe wie „Handlung" oder „Handelnde" haben unseres Erachtens Einfluss auf die Bestimmung des Intelligenzbegriffes und auf die Problemlösung (s. 5).

4. Pilotexperimente mit ChatGPT

4.1 Zum Stand der Forschung

Seit der Entstehung von ChatGPT im November 2022 sind mehrere Untersuchungen veröffentlicht worden, die sich mit der Umsetzung dieses Chatbots im Bereich der Lexikographie befassen. Einige dieser Studien (vgl. de Schryver und Joffe 2023; Rundell 2023; Jakubíček und Rundell 2023) vertreten die Ansicht, dass eine menschliche Bewertung und Überprüfung bei der Aufführung von lexikographischen Tätigkeiten immer noch erforderlich seien, obschon Maschinen die harte Arbeit — nämlich die Programmierung und Formatierung lexikographischer Daten — bereits übernehmen können. Andere Untersuchungen (vgl. Nichols 2023; Barrett 2023, zit. nach de Schryver 2023) deuten darauf hin, dass ChatGPT Informationen bieten könnte, die nicht vollkommen zuverlässig sind und Verbesserungen erfordern, und dass das System zum Teil „halluziniere".

Aus quantitativer Sicht (vgl. Phoodai und Rikk 2023) vermag ChatGPT jedoch eine bessere Leistung vorzuzeigen, denn das System ist in erster Linie dazu trainiert, Textlücken zu erfüllen.

Die erwähnten Arbeiten antworten teilweise auf die Fragen, die in der Einführung (s. 1) gestellt wurden. Außerdem kann zum einen festgestellt werden, dass lexikographische Kenntnisse immer noch von Vorteil sind, weil ExpertInnen

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1879 (Article)

im Bereich der Lexikographie diejenigen sind, die die Aufgabe übernehmen, die Antworten von ChatGPT auf lexikographisch-orientierte Fragen auszuwerten (vgl. de Schryver 2023; Alonso-Ramos 2023). In dieser Hinsicht ist anzumerken, dass der Mangel an Untersuchungen über die Handlungen der BenutzerInnen *in-actu* eklatant ist[4]. Zum anderen lässt sich feststellen, dass ChatGPT in der Lage ist, lexikographische Tätigkeiten auszuführen (vgl. de Schryver und Joffe 2023; Tran et al. 2023), wenn auch die Präzision und die Qualität der Ergebnisse etwas zu wünschen übrig lassen.

Daraus lässt sich schließen, dass die Benutzenden zu bewerten haben, wie kohärent der Output der Maschine ist. Laien hätten noch Schwierigkeiten, ein Prompt genauer zu erstellen. Im Falle einer lexikographischen Anfrage müssen sie noch dazu über die notwendigen spezifischen Kenntnisse verfügen. Hingegen können ExpertInnen nicht nur zur Evaluation und zur akribischeren Prompterstellung beitragen, sondern auch zur Verbesserung des Outputs. Das knüpft an die Frage nach der Effizienz und Qualität bei der Generierung lexikographischer Einträge seitens ChatGPT an.

Als eigene Textsorte müssen Wörterbuchartikel[5] — wie jeder Text — Kohärenz und Kohäsion aufweisen. Das bedeutet, dass der Output von ChatGPT einen gewissen Grad an inhaltlicher Verbindung und an textueller Organisation aufweisen muss. Wiegand (1996) erklärt, dass Wörterbuchartikel Teil des lexikographischen Textes sind und sich wiederum in Textteile (wie z.B. Lemmazeichengestaltangabe oder Bedeutungsangabe) segmentieren lassen. Davon ausgehend stellen wir die Frage, inwiefern ChatGPT als zur Erstellung von Texten trainiertes Sprachmodell einen lexikographischen Text mit einem bestimmten Grad an Kohärenz und Qualität (also Effizienz) verfassen kann. Im Folgenden wird das statistische Verfahren (s. 4.2) vorgestellt und die Outputs der Pilotexperimente (s. 4.3 und 4.4) angesichts ihres lexikographischen Wertes diskutiert.

4.2 Zur Methode und zur statistischen Auswertung

Zur Beantwortung der Frage, ob und inwiefern ChatGPT kohärente lexikographische Texte produziert, sind zwei ChatGPT-Sessions für jede der beiden Sprachen durchgeführt worden. Um die Ergebnisse quantitativ und qualitativ vergleichen zu können, haben wir nur ein Prompt ausgewählt, denn wir streben nicht an, das System zu trainieren oder den Prozess bis zur Erstellung eines zufriedenstellenden endgültigen Wörterbuchartikels zu evaluieren. Das vorgeschlagene Verfahren hat noch keine Aufmerksamkeit in der Fachliteratur gefunden, es entspricht aber der Mehrzahl der möglichen Verwendungen dieses Chatbots (s. Abbildung 1).

Die Auswahl des Deutschen und des Galicischen als Arbeitssprachen zielt darauf ab, andere Sprachen als das Englische zu prüfen, denn die meisten dem System zugrundeliegenden Daten, mehr als 90% (vgl. Rundell 2023), sind

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1879 (Article)

lediglich auf Englisch verfügbar. Es wird dadurch angestrebt, Einsichten in die Datenbasis zu gewinnen, auf der das Sprachmodell ChatGPT trainiert wurde.

Analysiert wird konkret die ChatGPT-Antwort auf den Prompt „Erstelle einen Wörterbuchartikel für L". Im galicischen Experiment musste zu dem Prompt hinzugefügt werden, dass das Ergebnis auf Galicisch erfasst werden sollte („Crea un artigo lexicográfico para L en galego"), weil sonst der Output oft auf Portugiesisch oder eben auf Spanisch erfolgt.

Die Abfragen auf Deutsch wurden für folgende Lemma formuliert: *abgeben*, *abholen*, *abschließen*, *anbieten* und *anmachen*. Es handelt sich um die fünf ersten trennbaren transitiven Verben der Goethe-Wortschatzliste zu A2. Für die Analyse der Ergebnisse auf Galicisch wurden die fünf im CORGA-Korpus häufigsten transitiven Verben ausgewählt: *facer*, *dicir*, *dar*, *saber* und *querer*.

Die in dieser Studie angewandten statistischen Methoden basieren auf der Verarbeitung von *n*-Grammen[6]. Zum quantitativen Vergleich der Outputs der zwei ChatGPT-Sessions werden zwei Metriken[7] umgesetzt: *Jaccard* und *Dice*[8]. Zum Vergleich der Outputs von der ersten Session in ChatGPT mit dem jeweiligen Wörterbuchartikel aus Referenzwörterbüchern wird eine dritte Metrik –*ROUGE-N*– herangezogen. Als Referenzwörterbücher gelten hier für das deutschsprachige Experiment der DUDEN und für das Experiment mit dem Galicischen das DRAG. Im Folgenden werden die Metriken sowie die erzielten Ergebnisse angeführt:

— Der *Jaccard*-Ähnlichkeitskoeffizient dient als Metrik zur Bewertung der Ähnlichkeit zwischen zwei Sets von *n*-Grammen und wird als Verhältnis der Anzahl gemeinsamer *n*-Gramme zur Gesamtzahl der Elemente beider Sets definiert. Der Text, der als Input gegeben wird, muss durch Leerzeichen tokenisiert werden und die Tokens werden in Sets umgewandelt, welche die statistische Berechnung ermöglichen. Der *Jaccard*-Koeffizient ergibt sich aus der Bestimmung von Schnittmengen und der Vereinigung der Sets. Der *Jaccard*-Koeffizient wird hier als Prozentsatz ausgedrückt.

— Der *Dice*-Koeffizient dient ebenso als Metrik zur Messung der Ähnlichkeit zwischen zwei Mengen. Zunächst werden die Texte durch Leerzeichen tokenisiert und in Vektoren von Wörtern umgewandelt. Diese Wörter werden dann in Sets transformiert, wodurch doppelte Wörter entfernt werden. Der *Dice*-Koeffizient wird berechnet, indem man die Anzahl der gemeinsamen Elemente verdoppelt und durch die Gesamtanzahl der Elemente beider Sets teilt. Der resultierende Koeffizient liegt zwischen 0 und 1, wobei 0 keine Ähnlichkeit und 1 vollständige Ähnlichkeit bedeutet. Das Ergebnis wird schließlich als Prozentsatz ausgedrückt.

Tabelle 1 gibt einen umfassenden Einblick in die statistischen Ergebnisse beider Verfahren. Die Prozentwerte verdeutlichen, in welchem Ausmaß ein Ähnlichkeitsverhalten zwischen den Outputs der beiden ChatGPT-Sitzungen festzustellen ist. Jede Zeile in der Tabelle bezieht sich auf die Wörterbuchartikel, die ChatGPT in den beiden Sitzungen für jedes ausgewählte Lemma generiert hat.

Als Beispiel für eine angemessene Interpretation der Daten lässt sich beobachten, dass im Fall des Lemmas *abgeben* eine Ähnlichkeit von 15,67% (*Jacard*) und 27,09% (*Dice*) zwischen dem Wörterbuchartikel der Sitzung 1 und dem der Sitzung 2 besteht. Außerdem wird in der letzten Zeile der Durchschnitt der Ähnlichkeit für jede Metrik und für jede Sprache angegeben.

	Pilotexperiment: Deutsch			**Pilotexperiment: Galicisch**		
	Jaccard	*Dice*		*Jaccard*	*Dice*	
abgeben	15,67%	27,09%		14,21%	24,88%	*facer*
abholen	22,40%	36,61%		14,01%	24,59%	*dicir*
abschließen	20,79%	34,42%		12,89%	22,83%	*dar*
anbieten	20,39%	33,86%		12,50%	22,22%	*saber*
anmachen	19,30%	32,35%		13,19%	23,3%	*querer*
	19,71%	**32,87%**	**Durchschnitt**	**13,36%**	**23,56%**	

Tabelle 1: Ähnlichkeitsverhalten zwischen den Outputs der zwei ChatGPT-Sitzungen

Die Ergebnisse weisen eine konsistente Tendenz auf, bei der der *Jaccard*-Koeffizient stets niedrigere Werte im Vergleich zum *Dice*-Koeffizienten anzeigt. Daraus ergibt sich, dass die Distanz zwischen den *Jaccard*- und *Dice*-Werten konstant oder zumindest ähnlich ist. Die konsistente Beziehung zwischen den beiden Koeffizienten bedeutet, dass die statistischen Verfahren angemessen angewandt worden sind. Auf der Grundlage des Pilotexperiments mit der deutschen Sprache stellt sich heraus, dass die erzielten Prozentsätze höher ausfallen, und dies veranschaulichen bereits die durchschnittlichen Ergebnisse: Beim Pilotexperiment mit dem Deutschen liegen die Werte bei 19,71% (*Jaccard*) und 32,87% (*Dice*), während die Werte im Experiment mit dem Galicischen 13,36% (*Jaccard*) und 23,56% (*Dice*) betragen. Die höheren Ergebnisse könnten mit einer umfassenderen Datengrundlage zusammenhängen, die für eine präzisere Modellierung und Textproduktion sorgt.

Aus dieser quantitativen Analyse lässt sich ebenfalls folgern, dass der Effizienzgrad von ChatGPT gering ist. Dies ist auf mehrere Faktoren zurückzuführen. Ein entscheidender Faktor ist die mangelnde Konstanz im generierten

Output. Die Unbeständigkeit in den erzeugten Texten kann zu Inkonsistenzen in der Qualität führen und erschwert eine zuverlässige Leistungsbewertung (s. 5.3). Dies könnte auf Schwächen im Trainingsprozess oder in den Modellparametern hinweisen, die zur Gewährleistung einer stabilen und konsistenten Ausgabe eine Überprüfung und Anpassung erfordern. Dieses Verfahren ermöglicht dennoch Einblicke in die Trainingsdaten, wobei man feststellen kann, dass die Daten, mit denen ChatGPT trainiert wurde, über weniger lexikographischen Text bzw. lexikographischen Textsorten verfügen. Außerdem müssen Benutzende den vom ChatGPT generierten Output bewerten (können), was zu einem erhöhten zeitlichen Aufwand führt und was die Benutzerfreundlichkeit des Systems beeinträchtigen kann. Die quantitativen Ergebnisse deuten zusammenfassend bei den beiden analysierten Sprachen darauf hin, dass ChatGPT bei lexikographischen Anfragen nicht präzise genug ist.

Die qualitativen Ergebnisse der *Jaccard*- und *Dice*-Metriken sind nicht das einzige entscheidende Bewertungskriterium, denn neben diesem wird die *ROUGE-N*-Methode eingesetzt, die eine effektive quantitative Auswertung der automatischen Textgenerierung im Vergleich zu Referenztexten ermöglicht[9]. Die *ROUGE-N*-Methode bietet eine Evaluierung der Textübereinstimmung an, insbesondere im Kontext der automatischen Textzusammenfassung. Hierbei erfolgt eine Tokenisierung der Texte in *n*-Gramme, die als Sets repräsentiert werden, um Duplikate zu eliminieren. Nach der Berechnung mathematischer Algorithmen fungieren die resultierenden F1-Score-Werte als quantitatives Maß für die Qualität automatisch von ChatGPT generierter Texte im Vergleich zu lexikographischen Referenztexten (in diesem Fall dem DUDEN und dem DRAG entnommen). Die Berechnung erfolgt durch die Überlappung von Unigrammen zwischen den Texten, wobei die *ROUGE*-1-Metrik F1-Score für die Ähnlichkeit ermittelt wird. Dies dient als Kriterium zur Beurteilung der Qualität automatisch generierter Texte. Zur Berechnung der *n*-Gramme und der Metriken werden die gesamten Wörterbuchartikel des DUDEN und DRAG genommen[10].

Tabelle 2 bietet einen Gesamtüberblick über die statistischen Ergebnisse der Textübereinstimmung zwischen dem von ChatGPT generierten lexikographischen Text und den Wörterbuchartikeln aus den ausgewählten Referenzwörterbüchern.

Die Ergebnisse der *ROUGE-N*-Metrik zeigen, dass die Ähnlichkeit zwischen den generierten ChatGPT-Texten und den Referenztexten gering ist, wobei nahezu so gut wie keine Übereinstimmung festzustellen ist. Der durchschnittliche Wert bei *ROUGE-N* für das Deutsche ist eine Übereinstimmung zwischen dem Wörterbuchartikel aus ChatGPT und dem aus Duden von 5,6%; im Falle des Galicischen beträgt sie 6,46%. Diese niedrigen *ROUGE-N*-Werte könnten mit Schwächen im Modelltraining, in den Daten oder in den Parametern zusammenhängen, die die Fähigkeit des Systems beeinträchtigen, relevante lexikographische Informationen adäquat zu extrahieren und wiederzugeben.

	Pilotexperiment: Deutsch		**Pilotexperiment: Galicisch**	
	ROUGE-N		*ROUGE-N*	
abgeben	4,95%		5,27%	*facer*
abholen	6,27%		7,02%	*dicir*
abschließen	4,72%		4,76%	*dar*
anbieten	6,90%		6,67%	*saber*
anmachen	5,17%		8,58%	*querer*
	5,60%	**Durchschnitt**	**6,46%**	

Tabelle 2: Übereinstimmung zwischen dem generierten lexikographischen ChatGPT-Text und den Wörterbuchartikeln aus Referenzwörterbüchern

Insgesamt lässt sich festhalten, dass die *Jaccard-*, *Dice-* und *ROUGE-N*-Werten niedrig sind und darüber hinaus, dass das System aus quantitativer Sicht nicht effizient ist (s. 5.4). Dies scheint auf eine potenzielle Begrenzung des lexikographischen Wissens des Systems hinzuweisen, da die niedrigen Prozentwerte auf eine geringe Übereinstimmung zwischen den generierten Texten und den erneut generierten Texten oder den Referenztexten hindeuten. Ein weiterer Aspekt, der in Betracht gezogen werden sollte, ist die Berücksichtigung der gesamten Mikrostruktur der Texte anstelle spezifischer Textabschnitte. Diese Herangehensweise ermöglicht eine umfassendere Bewertung der generierten Inhalte und kann zu niedrigeren prozentualen Ergebnissen im Vergleich zu dem Fall führen, in dem ein spezifischer Textteil (wie z.B. die Bedeutungsangabe oder die Beispielangabe) isoliert ausgewählt worden ist.

Weiterhin kann neben den quantitativen Ergebnissen eine qualitativ ausgerichtete Analyse der Textteile einen tieferen Einblick in den Effizienzgrad des Systems gewährleisten. Dazu setzen wir uns in 4.3 und 4.4 mit der Bewertung von spezifischen Textteilen und Aspekten der generierten Texte wie Kohärenz oder Relevanz auseinander.

4.3 ChatGPT als Verfasser von Wörterbuchartikeln im Deutschen

Zur Auswertung des Inhalts und der Outputs[11] achten wir auf die unterschiedlichen Textteilen — lexikographischen Angaben — in der Antwort von ChatGPT. Konkret setzen wir uns mit der Analyse der Angaben auseinander, die nach

Engelberg und Lemnitzer (2009: 135) wie folgt definiert werden: „die funktionalen Textsegmente, die in einer Angabebeziehung zu bestimmten Elementen [...] stehen und die es dem Benutzer ermöglichen sollen, aus ihnen bestimmte Informationen über den Wörterbuchgegenstand zu gewinnen."

Die hier diskutierten Angabeklassen beziehen sich grundsätzlich auf die inhaltliche Form der ausgewählten sprachlichen Einheiten. Die Textsegmente, deren Funktion darin besteht, die innere Zugriffsstruktur des lexikographischen Werkes deutlicher zu konzipieren, werden ausgeklammert. Die Ergebnisse lauten wie folgt:

(i) Lemmazeichengestaltangabe: Diese Angabe kommt in den zwei Sessions in derselben Form vor („Wörterbuchartikel: *L*"). Anzumerken ist dennoch, dass in der ersten ChatGPT-Session diese Angabe größer gedruckt wird.

(ii) Wortartangabe: Diese Angabe tritt in den zwei Sessions auf und ist immer richtig vorgegeben, denn die ausgewählten Wörter gehören zur Wortart der Verben. In der ersten Session wird diese Angabe fettgedruckt und in der zweiten Session ist die Wortartangabe kursiv hervorgehoben.

(iii) Ausspracheangabe: Diese Angabe wird nur in der zweiten Session präsentiert und erfolgt durch die Verwendung der IPA-Transkription, die zudem mit dem entsprechenden Akzent versehen wird. Die Darstellung der phonetischen Transkription ist richtig. Die Tatsache, dass diese Angabe nicht in jeder Session vorgestellt wird, lässt sich als eine der Schwächen des Systems verstehen.

(iv) Bedeutungsangabe: Die Bedeutungen und die verschiedenen Lesarten der ausgewählten Lemmata sind in den zwei Sessions vorhanden, aber werden in unterschiedlicher Form präsentiert. In der ersten Session werden zunächst sehr allgemeine Informationen angeboten: „Das Verb L hat verschiedene Bedeutungen, abhängig vom Kontext, in dem es verwendet wird". Erst nach dieser wenig aussagekräftigen Paraphrase werden die unterschiedlichen Lesarten aufgelistet, aber die Formulierungen bedürfen einer menschlichen Überprüfung, da diese gelegentlich wenig semantischbasiert sind. Folgendes Beispiel veranschaulicht dies: die dritte Lesart des Verbs *abholen* lautet „jemandem von einem Ort abholen".

In der zweiten Session wird keine allgemeine Definition vermittelt und die Lesarten werden durchnummeriert und vorgestellt. Die Formulierungen der Definitionen sind auch Kritik bedürftig, weil die Semantik hier wieder im Hintergrund steht. Ein Beispiel dafür ist die fünfte Lesart des Verbs *abschließen*: „etwas beenden, indem man es abschließt [...]". Im Fall der zweiten Session kommt als letzter Textteil ein Abschnitt vor, der „Verwendung" genannt wird und darunter fallen allgemeine Erklärungen hinsichtlich der Bedeutung der Lemmata: „Das Verb L wird in verschiedenen Kontexten verwendet [...]. Es findet Anwendung im Alltag, in formellen und informellen Gesprächen, im Handel, im Geschäftsleben, im Tourismus und vielen anderen Bereichen des täglichen Lebens". Solche allgemeinen Erläuterungen sind auch in der ersten Session unter

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1879 (Article)

der Überschrift „zusätzliche Informationen" vorhanden: „Das Verb L ist in der deutschen Sprache allgemein gebräuchlich und vielseitig einsetzbar".

(v) Flexionsangabe: Diese Angabe wird ausschließlich während der ersten Session erstellt. Dabei liegt der Fokus auf der Darstellung des Flexionsparadigmas der Verben im Präsens und Präteritum. In Wörterbüchern wird normalerweise das Flexionsparadigma nicht in seiner vollständigen Form angegeben, sondern lediglich durch drei Formen repräsentiert: Präsens (3. Person Singular), Präteritum (3. Person Singular), Partizip Perfekt.

(vi) Beispielangabe: Beispielsätze, die die Verwendung der Lemmata veranschaulichen, werden in den zwei Sessions vorgelegt, aber ihre Darstellung weicht in den beiden Sessions voneinander ab. In der ersten Session werden Beispielsätze unter der Überschrift „Verwendungsbeispiele" angeboten, aber sie werden nicht einer bestimmten Lesart zugeordnet. Bei jeder Lesart ist auch jeweils ein Beispielsatz vorhanden. In der zweiten Session hingegen tritt unter jeder Lesart ein Beispielsatz auf, der das kombinatorische Verhalten des Lemmas darstellt. Die Generierung dieser Beispielsätze ist oft wenig hilfreich und erklärend, da die Kombinatorik der Lemmata nicht deutlich dargestellt wird. Es kommen Beispielsätze wie folgende vor: „Sie hat sich in ihn verliebt und macht ihm Avancen" (Session 1, Le20mma: *anmachen*)[12] oder „Der Klempner muss die Rohre richtig abschließen, um Lecks zu vermeiden". (Session 2, Lemma: *abschließen*).

(vii) Synonymen- und Antonymenangaben: In beiden Sessions werden Synonyme und Antonyme zu den entsprechenden Lemmata angegeben, aber es erscheinen meist unterschiedliche Synonyme und Antonyme bei jeder Session, und deren Anzahl ist in der zweiten Session höher als in der ersten. Sogar Funktionsverbgefüge erscheinen als Synonyme.

(viii) Wortfamilienangabe: In der zweiten Session kommt ein Textteil vor, der „verwandte Begriffe" genannt wird. Dort werden immer zwei Substantive angegeben, die von dem entsprechenden Verb abgeleitet sind. Das erste Nomen verweist immer auf die durch das Verb beschriebene Handlung („Abgabe", „Abholung", „Abschluss" usw.), während sich das zweite Substantiv auf eine Person bezieht, die die Handlung vollzieht („Abgebender", „Abholer", „Anbieter" usw.).

(ix) Etymologische Angabe: Diese Informationen werden in den zwei Sessions geliefert, aber die vermittelten Informationen sind unterschiedlich. Als Beispiel dient der Fall des Verbs *abgeben*, das der ersten Session zufolge aus dem Präfix *ab* und dem althochdeutschen *gēben* stammt und das laut der zweiten Session dem mittelhochdeutschen *abegeben* entstammt. In der ersten Session wird der Ursprung der Verben auf das Althochdeutsche zurückgeführt, wogegen der Gebrauch derselben Lemmata laut der zweiten Session erst ab dem 16. und 17. Jahrhundert dokumentiert wurde.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1879 (Article)

Insgesamt lässt sich schließen, dass die angebotenen Informationen nicht konstant sind und dass sie sich teilweise widersprechen. Das deutet auf einen geringen Effizienzgrad des Systems hin, denn zur Erschließung sprachlicher Kenntnisse dank des generierten lexikographischen Texts ist eine Überprüfung in anderen Quellen erforderlich. Ein Vergleich mit den entsprechenden Wörterbuchartikeln im DUDEN weist ebenso darauf hin, dass die Mikrostruktur der lexikographischen Artikel im Referenzwörterbuch fixiert ist und dass dieselben Textteile in derselben Reihenfolge auftreten. ChatGPT bietet Informationen an, die zum Teil von denen im DUDEN abweichen.

4.4 ChatGPT als Verfasser von Wörterbuchartikeln im Galicischen

Kommen wir nun zum Experiment mit Galicisch als Metasprache[13]. Um es vorwegzunehmen: Das System von ChatGPT betrachtet das Galicische nicht immer als eigenständige Sprache, denn Textteile in der Antwort werden meist auf Portugiesisch und zum Teil auch auf Spanisch verfasst.

Auffällig beim Pilotexperiment mit Galicisch ist, dass im Output immer eine Einführung und eine Schlussfolgerung stehen. Offenbar hat ChatGPT das Wort „Artikel" hier im Sinne von ‚Aufsatz' verstanden, und deswegen neigt das System dazu, die Informationen ohne Kondensierung anzuordnen und mit einer gewissen Struktur vorzustellen. In den zwei Sessions kommen zum Beispiel Einführungen wie die folgende zustande:

> A palabra *L* é unha das palabras máis poderosas e frecuentes na lingua galega. Co seu uso tan común, é esencial comprender a súa ampla gama de significados e usos. Neste artigo lexicográfico, exploraremos a palabra L en galego, analizando as súas múltiples acepcións, construcións gramaticais e exemplos de uso.[14] (Session 1)

Nach der Beschreibung der Textteile, auf die wir gleich eingehen werden, kommen Schlussfolgerungen wie die folgende vor:

> O verbo L é um elemento central na língua galega devido à sua versatilidade e importância na comunicação. Sua conjugação e usos variados tornam-no uma ferramenta essencial para expressar uma ampla gama de ações, causas e estados de coisas. É uma palavra que transcende a gramática e se torna uma parte integrante da vida cotidiana para os falantes galegos, contribuindo para a riqueza e diversidade desta língua.[15] (Session 2)

Folgende mikrostrukturelle Textteile treten in den generierten galicischen Wörterbuchartikeln von ChatGPT vor:

(i) Lemmazeichengestaltangabe: Diese Angabe kommt in den zwei Sessions in Form einer kreativen Überschrift vor. In der Session 1 werden die Wörterbuchartikel mit einem Titel versehen (übersetzt: „*L* – eine lexikographische Analyse auf Galicisch"), während die Überschriften in der zweiten Session eher kreativ und attraktiv sind (übersetzt: „*L* – das vielseitige Verb der galicischen Sprache“).

(ii) Bedeutungsangabe: In der ersten Session werden die Lesarten der Lemmata durchnummeriert und getrennt angegeben. Die erste Lesart wird „primäre Bedeutung“ genannt. Die Definitionen beginnen immer mit der Formulierung „Bedeutung von…“, was in der lexikographischen Praxis nicht üblich ist. In der zweiten Session wird eine einzige Definition angegeben, die normalerweise auf mehrere Lesarten Bezug nimmt. Bei der Erfassung der Definitionen werden häufig wenig aussagekräftige Informationen geliefert: „É um dos verbos mais comuns e essenciais na língua galega, desempenhando um papel fundamental na comunicação e transmissão de informações“[16]. Hinzu sollte noch ergänzt werden, dass die Lesarten der Lemmata in der Session 2 unter der Überschrift „häufige Verwendungen“ verzeichnet werden.

(iii) Flexionsangabe: Diese Angabe kommt in beiden Sessions vor; in der ersten Session wird erklärt, dass die Verben in allen Tempora konjugiert werden können. Beispiele von vier Tempora (Präsens, zwei Formen des Präteritums, Futur) werden geboten. In der zweiten Session hingegen werden die Verben im Präsens vollständig konjugiert.

(iv) Idiomangabe: In der ersten Session ist ein Textteil vorhanden, in dem idiomatische Redewendungen aufgelistet werden. Eine menschliche Überprüfung zeigt aber, dass nicht alle als Idiome aufgefasst werden können. Beispiele davon sind „saber **a algo**“ (‚**nach etwas** schmecken‘) oder „dicir unha mentira piadosa“ (‚eine Notlüge erzählen‘).

(v) Beispielangabe: Es wird kein Textteil erzeugt, in dem Beispielsätze systematisch den Gebrauch eines Lemmas beschreiben. Auf jede Lesart folgt aber in der ersten Session eine Liste von „Beispielen“, die teilweise die Verwendung der lexikalischen Einheiten veranschaulichen. Einige dieser Beispiele fungieren unserer Ansicht nach als Kollokationen in der galicischen Sprache: „dar un paseo“ (‚einen Spaziergang machen‘) oder „dar consello“ (‚Ratschläge geben‘).

(vi) Angabe der diatopischen Variation: In der zweiten Session werden Informationen hinsichtlich der diatopischen Variation gegeben, aber es handelt sich um keine genauen Informationen, die Bezug auf die Variation eines bestimmten Lemmas nehmen. Dieser Teil hält fest, wie varietätenreich die galicische Sprache ist.

Vergleicht man diese Informationen mit der Struktur des DRAG, stellt man fest, dass beim Output keine Systematisierung hinsichtlich der Anordnung, Gliederung und Beschreibung der Textteile existiert. Ähnlich wie im Fall des deutschen Experiments fehlt es an konstanten Informationen, was zu einem mangelnden Effizienzgrad führt. Die Unbeständigkeit oder das Fehlen von konstanten bzw. kohärenten[17] Daten beeinträchtigt maßgeblich die Wirksamkeit des Systems, da eine zuverlässige Grundlage für Entscheidungen in der lexikographischen Beschreibung fehlt. Dieser Mangel an Stabilität kann potenziell zu Unsicherheiten und ineffizienten Abläufen auf der Seite der BenutzerInnen führen.

5. Gesamtanalyse der Ergebnisse

Im letzten Abschnitt sollen die Hauptergebnissen der Untersuchung ausgeführt werden, die unausgesprochen auf die Unterscheidung zwischen der Lexikographie und den Sprachmodellen bzw. der künstlichen Intelligenz hindeuten.

5.1 Handlung und Handelnde

Einige Inkonsistenzen bei den Ergebnissen von ChatGPT lassen sich durch die Handlung und die Handelnden erläutern. Beim Erstellungsprozess eines lexikographischen Werkes ist der/die Lexikograph/in (oder ein Laie, wie bei einigen kollaborativen Werken) als das denkende Agens zu verstehen. Was die Benutzung betrifft, ist bei der Lexikographie der primäre Adressat der Handlung ein menschlich denkendes Wesen, der sekundäre Adressat eine Maschine (s. Abbildung 2). Im Falle der KI konzipiert ein/e Entwickler/in von Computersystemen und Software ein intelligentes System. Bei der KI ist unseres Erachtens der primäre Adressat ein nicht denkendes Wesen — eine Maschine —, sekundäre Adressaten sind wieder Maschinen, aber auch Menschen. Unangesprochen agieren die primären und sekundären Adressaten in Hinblick auf die Denkprozesse anders (wenn die Rede überhaupt in allen Fällen von Denkprozessen sein kann).

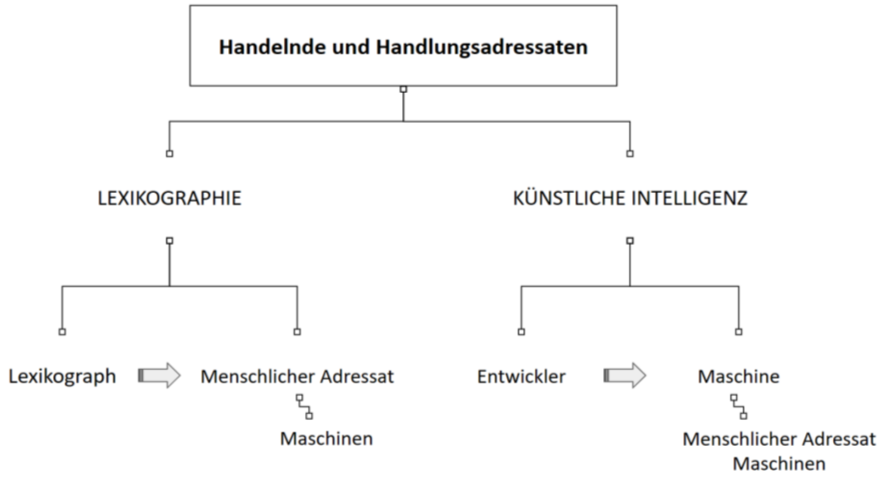


Abbildung 2: Handelnde und Adressaten in der Lexikographie und der KI

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1879 (Article)

Zieht man die Unterscheidung zwischen Computerlexikographie und computergestützten Lexikographie in Betracht, ist das vorangehende Schema wie folgt weiter zu ergänzen:

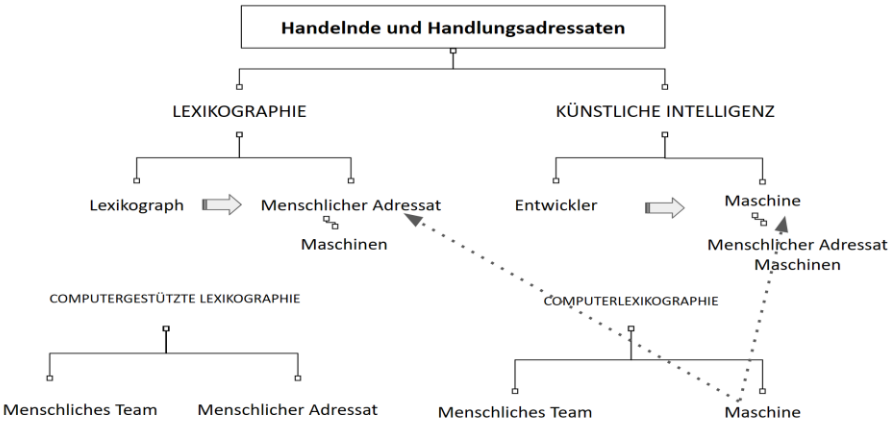


Abbildung 3: Handelnde und Adressaten in der Computerlexikographie, der computergestützten Lexikographie und der KI

Sowohl LexikographInnen[18] als auch menschliche BenutzerInnen[19] von lexikographischen Ressourcen möchten Antwort auf eine Frage finden. In diesem Zusammenhang stellen das lexikographische Team sowie die menschlichen BenutzerInnen Hypothesen auf. Im Falle der KI wird die einzige kognitive Hypothesenbildung vom Entwicklungsteam und von sekundären menschlichen BenutzerInnen hergestellt. Maschinen per se bilden keine Hypothese: Intelligente Systeme verarbeiten auf der Grundlage von Algorithmen Daten und in Anlehnung daran bieten sie ein Output. Kreativ — im Sinne von kognitiven Prozessen — sind sie aber nicht.

Dies lässt sich bei dem Output in den ChatGPT-Sessions (s. 4.3 und 4.4) beobachten. Das System aktiviert bei der Beantwortung des Prompts einen konkreten Wissensbereich. Wenn nicht genügende Informationen vorhanden sind, dann wendet ChatGPT Ähnlichkeitsprinzipien an; das erklärt, warum es in den galicischen Antworten auf das Portugiesische oder auf das Spanische zurückgreift. Diese Ergebnisse lassen sich angesichts der Angemessenheit hinsichtlich der Kohärenz, Kohäsion oder Textsortenspezifizität nur von einem menschlichen Agens evaluieren[20] (vgl. Effizienz in 5.4). Diese kognitive Evaluationshandlung erlaubt den Schluss, dass die von ChatGPT generierte Antwort im Galicischen hinsichtlich „Wörterbuchartikel" auf „Artikel" im Sinne von ‚Abhandlung, Aufsatz' zurückgeht (vgl. 4.4). Dies erklärt die ChatGPT-Vermittlung der Informationen in vollständigen Absätzen in wissenschaftlicher

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1879 (Article)

Form, was der lexikographischen Tradition eher widerspricht: In Wörterbuchartikeln wird jedem Textteil eine Funktion zugewiesen und die Information wird meist kondensiert.

5.2 Der genuine Zweck und der Intelligenzbegriff

Zur Auseinandersetzung mit dem Intelligenzbegriff sei an den genuinen Zweck eines Wörterbuches erinnert:

> Der genuine Zweck eines Wörterbuches besteht darin, dass es benutzt wird, um anhand lexikographischer Daten in den Teiltexten mit äußerer Zugriffsstruktur […] Informationen zu denjenigen Eigenschaftsausprägungen bei sprachlichen Ausdrücken zu erschließen, die zum jeweiligen Wörterbuchgegenstand gehören. (Wiegand 1998: 299)

Überträgt man diese Definition auf intelligente Systeme, kommt man zur Schlussfolgerung, dass der „genuine Zweck" eines intelligenten Systems darin besteht, anhand einer großen Datenmenge und nach einer Phase maschinellen Lernens Handlungen teilweise auszuführen, die Menschen bei unterschiedlichen Aktivitäten unterstützen können. Derartige Systeme haben aber keineswegs die Absicht, lexikographisches Wissen zu vermitteln, und ihre Benutzung als Nachschlagewerke gewährleistet nicht, dass die Benutzenden Einsichten in das sprachliche Verhalten einer gewissen lexikalischen Einheit oder in die Wörterbuchartikel als Textsorte gewinnen. Das lässt sich bei der Analyse der Ergebnisse in vorangehenden Abschnitten beobachten.

5.3 Daten: Präsentation, Systematisierung und Qualität

Die Pilotexperimente zeigen, dass eine Systematisierung im Output von ChatGPT fehlt und dass das System auf wenigen lexikographischen Texten trainiert wurde. Die Metriken (s. 4.2) weisen eine geringe Ähnlichkeit zwischen den Outputs in den zwei Sitzungen auf, die sich auch bei der qualitativen Analyse nachweisen lässt. Hinsichtlich dieser mangelnden Einheitlichkeit vertreten Jakubíček und Rundell (2023) die Ansicht, dass ChatGPT keine festgelegten Kriterien hat. Das lässt sich z.B. an den generierten Daten der mikrostrukturellen Textteile beobachten, indem die Lemmazeichengestaltangabe und die Bedeutungsangabe die einzigen regelmäßig vorhandenen Angaben sind. Es lässt sich vermuten, dass diese die einzigen beständigen lexikographischen Textteile in den Trainingsdaten vom ChatGPT darstellen. Dennoch sind die Output-Unterschiede in den beiden analysierten Sprachen nicht zu übersehen (vgl. 4.3 und 4.4). Dies geht unseres Erachtens darauf zurück, dass die meisten Daten im Chatbot auf Englisch sind (vgl. Jakubíček und Rundell 2023), dass das System *crosslingual* operiert und dass es auf wenigen Daten auf Galicisch

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1879 (Article)

beruht. Bezüglich der Datenqualität sei hier zu betonen, dass die Mehrheit der identifizierten Fehler auf der semantischen Ebene zu verorten ist.

5.4 Effizienzgrad und Lösungsverfahren

Die wenig systematisierten und begrenzt einheitlichen Outputs machen deutlich, dass das Verfahren zur Lösung lexikographischer Probleme bzw. Fragen bei ChatGPT nicht effizient ist (vgl. Radford et al. 2018): Je nach Kontext und bedingt durch das vorherige Gespräch mit den BenutzerInnen antwortet ChatGPT anders auf die Fragen. Das bedeutet dann, dass der Effizienzgrad intelligenter Systeme nicht nur nach der Datenmenge evaluiert werden kann (vgl. Phoodai und Rikk 2023).

Zur endgültigen Evaluation des Effizienzgrades muss außerdem die Rolle der Benutzenden in Erwägung gezogen werden. Benutzende/Handelnde agieren nicht passiv, denn sie wollen sich neue Informationen erschließen. Die denkenden Handelnden — sowohl LexikographInnen als auch BenutzerInnen — können die angebotenen Informationen auswerten und der Handlung entsprechend agieren. Unterscheidet man zwischen Laien und ExpertInnen[21], wird der Effizienzgrad des Problemlösungsverfahrens davon abhängen, wie kundig ein bestimmter Benutzer oder eine bestimmte Benutzerin ist. Folglich kann ein/e erfahrene/r Benutzer/in imstande sein, sich die Informationen zu erschließen, die vom intelligenten System geliefert werden. Hingegen kann ein/e unkundige/r Benutzer oder Benutzerin — zum Beispiel ein/e Sprachlernende/r mit begrenzter Sprachkompetenz– nicht validieren, ob die Informationen den Tatsachen entsprechen oder nicht.

6. Zusammenfassung

Neben der quantitativen und qualitativen Evaluation befasst sich der Aufsatz auch mit Gemeinsamkeiten und Unterschieden zwischen der Lexikographie und der künstlichen Intelligenz, somit werden nicht nur der Effizienzbegriff im Sinne des Wirkungsgrades und der Dateninterpretation, sondern auch der Intelligenzbegriff, die Handlung und die Handelnden bei kognitiven Prozessen diskutiert.

Die Analyse von ChatGPT als hilfeleistendem Werkzeug bei einem lexikographischen Projekt ist nicht das Ziel der Untersuchung. Aus diesem Grund ist das Chatbot mit unterschiedlichen Prompts nicht traniert worden. Offene Fragen für künftige Studien bestehen in der Interaktion zwischen der Lexikographie und ChatGPT, z.B. inwiefern die Lexikographie zwecks der Überarbeitung und Anpassung generierter Wörterbuchartikel von diesem Chatbot profitieren kann. Es ist festgestellt worden, dass ChatGPT über Daten bezüglich der Textsorte „Wörterbuchartikel" im Deutschen und im Galicischen nicht verfügt. Die Datenverfügbarkeit bei anderen Sprachen bedarf einer weiteren Untersuchung.

Insgesamt lässt sich festhalten, dass zurzeit ChatGPT die lexikographische

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1879 (Article)

Arbeit nicht ersetzen kann: ChatGPT kann nicht als Garant für das kulturelle und sprachliche Kulturerbe gelten, es gilt nicht als ein kultureller, sozialer, normativ-linguistischer und sogar rechtlicher Bezugspunkt.

Diese Studie zeigt, dass heutzutage die Lexikographie von der künstlichen Intelligenz nicht ersetzbar ist, denn die KI kann eine vollständige lexikographische Tätigkeit nicht ausführen. Das ist eben nicht ihr Ziel. Beide weisen unterschiedliche Aktoren, Handlungen, Forschungsgegenstände, Analyseverfahren, Ziele und Ergebnisse auf.

Danksagung

Die Ergebnisse dieser Forschung stehen im Zusammenhang mit dem Untersuchungsvorhaben „Automatic and sustainable multilingual semantic tagger — ESMAS-ES^{+}“, gefördert vom spanischen Ministerium für Wissenschaft und Innovation/Staatliche Forschungsagentur/FEDER „A way to make Europe” (Projekt PID2022-137170OB-I00).

Die Durchführung dieser Studie war z.T. dank des Programms *Formación de Profesorado Universitario* (Förderkennzeichen FPU21/00188) vom spanischen Ministerium für Universitäten möglich.

Endnoten

1. ChatGPT ist ein von OpenAI entwickelter Chatbot, der Antworten in natürlichen Sprachen generieren kann. In dieser Studie arbeiten wir lediglich mit der kostenfreien Version 3.5.
2. Im Bereich der KI und der natürlichen Sprachverarbeitung spricht man von Prompt, um eine Eingabeaufforderung zu benennen, die ein Mensch an ein System — in diesem Fall an ChatGPT — richtet und die den Chatbot dazu veranlasst, eine Antwort zu generieren.
3. Weitere Beispiele sind Kabashi (2018) und Delli Bovi und Navigli (2017).
4. Eine Ausnahme bilden Müller-Spitzer et al. (2018). Sie beziehen sich aber auf die Verwendung von online Sprachressourcen, nicht konkret auf ChatGPT.
5. Im WLWF-4 (2020: 120) wird unter „Wörterbuchartikel“ das Folgende verstanden: „aus der geordneten Menge von Textelementen und/oder Textbausteinen bestehender Textteil des Wörterverzeichnisses, für den das Lemma obligatorisch ist und in dem mindestens eine Eigenschaft des Lemmazeichens beschrieben oder angeführt wird.“
6. *n*-Gramme sind aufeinanderfolgende Wortsequenzen, wobei *n* die Fenstergröße der Sequenzen kodiert. Es bestehen bereits lexikalische Ressourcen und Systeme, die linguistische Daten auf der Grundlage von *n*-Grammen verarbeiten und darstellen (vgl. Wolfer et al. 2023).
7. Diese Metriken werden auch zur Analyse der Ähnlichkeit von zwei Texten herangezogen (vgl. Stefanovič et al. 2019).
8. Auf Anfrage stellen wir den R-Code zur Berechnung der Ähnlichkeitskoeffizienten *Jaccard*, *Dice* und *ROUGE-N* zur Verfügung.
9. Diese Methode wurde auch in anderen Studien lexikographischer Art (vgl. Phoodai und Rikk 2023) angewandt.

10. Für unsere Belange muss geprüft werden, in welchem Ausmaß eine Ähnlichkeitstendenz zwischen den gesamten Wörterbuchartikeln festzustellen ist. Nimmt man spezifische Textteile (wie etwa die Definition), könnte das Verhalten auf einen höheren Grad von Ähnlichkeit verweisen.
11. Die zugrundeliegenden ChatGPT-Sessions sind unter folgenden Links abrufbar:
    (a) Session 1: https://chat.openai.com/share/9c89990d-7e65-4173-9578-d8e613b3a880
    (b) Session 2: https://chat.openai.com/share/c087e4f4-2069-4605-aead-268c9ea1cd47.
12. Es ist hier erwähnenswert, dass im Beispiel nicht einmal das beschriebene Lemma vorkommt.
13. Die ChatGPT-Sessions mit Galicisch als Metasprache sind unter folgenden Links abrufbar:
    (a) Session 1: https://chat.openai.com/share/0ec3c864-0508-4cef-b8d5-c93c43e05c3e,
    (b) Session 2: https://chat.openai.com/share/55313ca3-90dc-4fcf-9dbb-36721b03ee7e.
14. Übersetzung von uns: „Das Wort L ist eines der mächtigsten und häufigsten Wörter in der galicischen Sprache. Da es so häufig verwendet wird, ist es wichtig, seine vielfältigen Bedeutungen und Verwendungen zu verstehen. In diesem lexikographischen Artikel werden wir das Wort L im Galicischen untersuchen und seine vielfältigen Bedeutungen, grammatikalischen Konstruktionen und Verwendungsbeispiele analysieren."
15. Übersetzung von uns: „Das Verb L ist aufgrund seiner Vielseitigkeit und Bedeutung in der Kommunikation ein zentrales Element der galicischen Sprache. Seine Konjugation und sein vielfältiger Gebrauch machen es zu einem unverzichtbaren Werkzeug, um eine Vielzahl von Handlungen, Ursachen und Zuständen auszudrücken. Es ist ein Wort, das über die Grammatik hinausgeht und für die Sprecher des Galicischen zu einem festen Bestandteil des täglichen Lebens wird, was zum Reichtum und zur Vielfalt dieser Sprache beiträgt."
16. Übersetzung von uns: „Es ist eines der gebräuchlichsten und wichtigsten Verben in der galicischen Sprache und spielt eine grundlegende Rolle in der Kommunikation und der Übermittlung von Informationen."
17. Damit ist gemeint, dass das ChatGPT-System in jeder Sitzung auf einen anderen Datenabschnitt zurückgreift, sodass die bereitgestellten Informationen von vorherigen abweichen (können).
18. Im Falle der LexikographInnen geht es darum, wie bestimmte Inhalte lexikographisch vermittelt werden können, wer der Adressat des Werkes sein wird, u.a. Dabei versucht man neben computerbezogenen Fragen im lexikographischen Prozess (Müller-Spitzer 2007: 17), auch weitere, „die darauf abzielen, diejenigen theoretischen, methodischen, terminologischen, historischen, dokumentarischen, didaktischen und kulturpädagogischen Fragen zu beantworten" (Müller-Spitzer 2007: 17).
19. Die BenutzerInnen machen unterschiedliche kognitive Schritte bei einer Suchanfrage wie z.B.: „Was möchte ich konkret nachschlagen?", „Wie gehe ich bei einer erfolglosen Suche vor?", u.a. Dabei stellen sie eine Hypothese oder Teilhypothesen im ganzen Verfahren auf (Domínguez Vázquez und Valcárcel Riveiro 2015; Müller-Spitzer et al. 2018).
20. Führt man dennoch die Suchanfrage in einem intelligenten System aus, wird man damit konfrontiert, dass man nicht nur Informationen aus den Daten extrahieren sollte, sondern man muss auch bewerten, inwiefern die gelieferte Information richtig ist. Benutzende sollten sich dessen bewusst sein, dass die Informationen, die sie erhalten, nicht unbedingt richtig sein müssen und dass sie zur Formulierung einer fundierten Antwort andere Ressourcen konsultieren sollten.

21. Unter die Gruppe der Laien fallen beispielsweise Sprachlernende oder SchülerInnen. Zur Gruppe der ExpertInnen gehören LexikographInnen, SpezialistInnen und zum Teil auch erfahrene BenutzerInnen.

Literaturverzeichnis

Alonso-Ramos, Margarita. 2023. El papel de ChatGPT como lexicógrafo. Garriga Escribano, Cecilio, Sandra Iglesia Martín, José Antonio Moreno Villanueva und Antoni Nomdedeu Rull (Hrsg.). 2023. *Lligams: Textos dedicats a Maria Bargalló Escrivà*: 15-27. Tarragona: Publicacions URV.

Barrett, Grant. 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*.

CORGA = Centro Ramón Piñeiro para a Investigación en Humanidades. 2022. *CORGA: Corpus de Referencia do Galego Actual*.
https://corpus.cirp.es/corga/. Letzter Zugriff: 04.01.2024.

Delli Bovi, Claudio und Roberto Navigli. 2017. Multilingual Semantic Dictionaries for Natural Language Processing: The Case of BabelNet. *World Scientific Encyclopedia with Semantic Computing and Robotic Intelligence*. *Semantic Computing*: 149-163.
doi: 10.1142/S2425038416300159

de Schryver, Gilles-Maurice. 2023. Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography* 36(4): 1-33.
doi: 10.1093/ijl/ecad021

de Schryver, Gilles-Maurice und David Joffe. 2023. The End of Lexicography, Welcome to the Machine: On How ChatGPT Can Already Take over All of the Dictionary Maker's Tasks. *20th CODH Seminar, Center for Open Data in the Humanities, Research Organization of Information and Systems, Tokyo, Japan, 27 February 2023.* Tokyo: National Institute of Informatics.
https://youtu.be/watch?v=mEorw0yefAs. Letzter Zugriff: 10.12.2023.

Domínguez Vázquez, María José und Carlos Valcárcel Riveiro. 2015. Hábitos de uso de los diccionarios entre los estudiantes universitarios europeos: ¿nuevas tendencias? Domínguez Vázquez, María José, Xavier Gómez Guinovart und Carlos Varcárcel Riveiro (Hrsg.). 2015. *Lexicografía de las lenguas románicas. II: Aproximaciones a la lexicografía moderna y contrastiva*: 165-190. Berlin/Boston: De Gruyter.
doi: 10.1515/9783110310337.165

Engelberg, Stefan und Lothar Lemnitzer. 2009. *Lexikographie und Wörterbuchbenutzung*. Tübingen: Stauffenburg.

Goethe Institut. 2016. *Goethe-Zertifikat A2: Wortliste*. München: Goethe-Institut.

Jakubíček, Miloš und Michael Rundell. 2023. The End of Lexicography? Can ChatGPT Outperform Current Tools for Post-editing Lexicography? Medved', Marek, Michal Měchura, Carole Tiberius, Iztok Kosem, Jelena Kallas, Miloš Jakubíček und S. Krek (Hrsg.). 2023. *Electronic Lexicography in the 21st Century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 Conference, Brno, 27–29 June 2023:* 518-533. Brno: Lexical Computing CZ s.r.o.

Kabashi, Besim. 2018. A Lexicon of Albanian for Natural Language Processing. Čibej, Jaka, Vojko Gorjanc, Iztok Kosem und Simon Krek (Hrsg.). 2018. *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts, 17–21 July 2018, Ljubljana, Slovenia:* 855-862. Ljubljana: Ljubljana University Press.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1879 (Article)

Kosem, Iztok, Robert Lew, Carolin Müller-Spitzer, Maria Ribeiro Silveira und Sascha Wolfer (Koord.). 2019. The Image of the Monolingual Dictionary Across Europe. Results of the European Survey of Dictionary Use and Culture. *International Journal of Lexicography* 32(1): 92-114.
doi: 10.1093/ijl/ecy022

Kouassi, Konan. 2022. Mensch-Maschine-Interaktion im lexikographischen Prozess zu lexikalischen Informationssystemen. Klosa-Kückelhaus, Annette, Stefan Engelberg, Christine Möhrs und Petra Storjohann (Hrsg.). 2022. *Dictionaries and Society. Proceedings of the XX EURALEX International Congress, 12–16 July 2022, Mannheim, Germany:* 172-180. Mannheim: IDS-Verlag.

Mainzer, Klaus. 2019. Einführung: Was ist KI? *Künstliche Intelligenz — Wann übernehmen die Maschinen?* Technik im Fokus: 1–6. Berlin/Heidelberg: Springer.
doi: 10.1007/978-3-662-58046-2_1

McCarthy, John. 2007. *What is AI? Basic Questions* (Interview with Prof. John McCarthy).
http://jmc.stanford.edu/artificial-intelligence/what-is-ai/index.html. Letzter Zugriff: 20.12.2024.

Mel'čuk, Igor. 1984. *Dictionnaire explicatif et combinatoire du français contemporain. Recherches lexico-sémantiques.* Amsterdam: John Benjamins.

Müller-Spitzer, Carolin. 2007. *Der lexikographische Prozess. Konzeption für die Modellierung der Datenbasis.* Studien zur Deutschen Sprache 42. Tübingen: Narr.

Müller-Spitzer, Carolin, María José Domínguez Vázquez, Martina Nied Curcio, Idalete Maria Silva Dias und Sascha Wolfer. 2018. Correct Hypotheses and Careful Reading Are Essential: Results of an Observational Study on Learners Using Online Language Resources. *Lexikos* 28: 287-315.
doi: 10.5788/28-1-1466

Müller-Spitzer, Carolin und Alexander Koplenig. 2015. Requisitos y expectativas de un buen diccionario online. Resultados de estudios empíricos en la investigación sobre el uso de diccionarios, con especial atención a los traductores. Domínguez Vázquez, María José, Xavier Gómez Guinovart und Carlos Valcárcel Riveiro (Hrsg.). 2015. *Lexicografía de las lenguas románicas. II: Aproximaciones a la lexicografía moderna y contrastiva*: 297-319. Berlin/Boston: De Gruyter.

Nallapati, Ramesh, Bowen Zhou, Cicero dos Santos, Çağlar Gu̇lçehre und Bing Xiang. 2016. Abstractive Text Summarization Using Sequence-to-sequence RNNs and Beyond. Riezler, Stefan und Yoav Goldberg (Hrsg.). 2016. *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL), Berlin, Germany, August, 11–12, 2016:* 280-290. Berlin: Association for Computational Linguistics.

Nichols, Wendalyn. 2023. Invisible Lexicographers, AI, and the Future of the Dictionary. *Paper presented at the eLex 2023 Conference, Brno, 27–29 June 2023: Electronic Lexicography in the 21st Century.*
https://www.youtube.com/watch?v=xYpwftj_QQI. Letzter Zugriff: 03.01.2024.

Phoodai, Chayanon und Richárd Rikk. 2023. Exploring the Capabilities of ChatGPT for Lexicographical Purposes: A Comparison with Oxford Advanced Learner's Dictionary within the Microstructural Framework. Medved', Marek, Michal Měchura, Carole Tiberius, Iztok Kosem, Jelena Kallas, Miloš Jakubíček und S. Krek (Hrsg.). 2023. *Electronic Lexicography in the 21st Century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 Conference, Brno, 27–29 June 2023*: 345-375. Brno: Lexical Computing CZ s.r.o.

Radford, Alec, Karthik Narasimhan, Tim Salimans und Ilya Sutskever. 2018. *Improving Language Understanding by Generative Pre-Training.*
https://openai.com/research/language-unsupervised. Letzter Zugriff: 15.12.2023.

Roemmele, Melissa. 2016. Writing Stories with Help from Recurrent Neural Networks. Schuurmans, Dale und Michael Wellman (Hrsg.). 2016. *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12–17 2016, Phoenix, Arizona USA (AAAI-16)* 30(1): 4311-4312. California: AAAI Press.

Rundell, Michael. 2023. Automating the Creation of Dictionaries: Are We Nearly There? *Asialex 2023 Proceedings. Lexicography, Artificial Intelligence, and Dictionary Users, 22–24 June 2023, Seoul, Korea:* 9-17. Seoul: Yonsei University.

Stefanovič, Pavel, Olga Kurasova und Rokas Štrimaitis. 2019. The N-Grams Based Text Similarity Detection Approach Using Self-Organizing Maps and Similarity Measures. *Applied Sciences* 9(9), 1870: 1-14.
doi: 10.3390/app9091870

Tran, Hanh Thi Hong, Vid Podpečan, Mateja Jemec Tomazin und Senja Pollak. 2023. Definition Extraction for Slovene: Patterns, Transformer Classifiers and ChatGPT. Medved, Marek, Michal Měchura, Carole Tiberius, Iztok Kosem, Jelena Kallas und Miloš Jakubíček (Hrsg.). 2023. *Electronic Lexicography in the 21st Century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 Conference, Brno, 27–29 June 2023*: 19-38. Brno: Lexical Computing s.r.o.

Wiegand, Herbert Ernst. 1983. Überlegungen zu einer Theorie der lexikographischen Sprachbeschreibung. Hyldgaard-Jensen, Karl und Arze Zettersten (Hrsg.). 1983. *Symposium zur Lexikographie. Symposium on Lexicography. Proceedings of the Symposium on Lexicography, September 1–2 1982, Universität Kopenhagen.* Germanistische Linguistik 5–6: 35-72. Hildesheim/New York: Georg Olms.

Wiegand, Herbert Ernst. 1996. A Theory of Lexicographic Texts: An Overview. *South African Journal of Linguistics* 14(4): 134-149.
doi: 10.1080/10118063.1996.9724061

Wiegand, Herbert Ernst. 1998. *Wörterbuchforschung: Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie.* Berlin: De Gruyter.

Wikipedia. Intelligenz: https://de.wikipedia.org/wiki/Intelligenz. Letzter Zugriff: 05.01.2024.

WLWF-3 = Wiegand, Herbert Ernst, Rufus H. Gouws, Matthias Kammerer, Michael Mann und Werner Wolski (Hrsg.). 2020. *Wörterbuch zur Lexikographie und Wörterbuchforschung. Band 3*. Berlin/Boston: De Gruyter.

WLWF-4 = Wiegand, Herbert Ernst, Rufus H. Gouws, Matthias Kammerer, Michael Mann und Werner Wolski (Hrsg.). 2020. *Wörterbuch zur Lexikographie und Wörterbuchforschung. Band 4*. Berlin/Boston: De Gruyter.

Wolfer, Sascha, Alexander Koplenig, Marc Kupietz und Carolin Müller-Spitzer. 2023. Introducing DeReKoGram: A Novel Frequency Dataset with Lemma and Part-of-Speech Information for German. *Data* 8(11): 1-10.
doi: 10.3390/data8110170

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1880 (Article)

Dictionary Use Training in Secondary School EFL Textbooks in Taiwan

Wai-on Law, *School of Modern Languages,
University of St Andrews, Scotland
(wai-on.law@dunelm.org.uk) (https://orcid.org/0000-0001-5887-573X)*

Abstract: As a rare study on English language textbook design for dictionary use training, this research examines four series of secondary school textbooks available on the Taiwan market. The content analysis method was adopted in finding out (1) how effectively the existing secondary English textbooks can help learners develop the necessary dictionary skills based on the guidelines from the government; and (2) how the existing textbooks could be improved to better meet learners' needs for dictionary skills training. The results show that none of the surveyed series follow the Curriculum Guidelines of the Ministry of Education (2018) regarding dictionary use training, although they all claim to have designed the book based on the government-set curriculum. Suggestions are made regarding how the present textbook designs could incorporate dictionary skills, with recommended resources. The study reveals the conspicuous neglect of dictionary use skills training in secondary school textbooks, and calls for similar review to be made in other countries to fully appreciate the (un)availability of dictionary use training in secondary schools. The study should provide useful information to relevant government authorities, dictionary compilers, textbook writers, and English language teachers and researchers alike for improving the situation.

Keywords: DICTIONARY USE TRAINING, TAIWAN, SECONDARY SCHOOL EFL TEXTBOOK, CONTENT ANALYSIS

Opsomming: Woordeboekgebruiksopleiding in sekondêre skool-EVT-handboeke in Taiwan. As minder gebruiklike studie van die ontwerp van Engelse taalhandboeke vir woordeboekgebruiksopleiding word in hierdie navorsing vier reekse sekondêre skoolhandboeke wat op die Taiwannese mark beskikbaar is, ondersoek. Die inhoudsanalisemetode is toegepas om te bepaal (1) hoe effektief die huidige sekondêre Engelse handboeke leerders kan help om die nodige woordeboekvaardighede gebaseer op die riglyne van die regering, te ontwikkel; en (2) hoe die huidige handboeke verbeter kan word om beter voorsiening te kan maak vir die leerders se behoeftes aan woordeboekvaardigheidsopleiding. Die resultate toon aan dat geeneen van die nagevorste reekse die Kurrikulumriglyne van die Ministerie van Onderwys (2018) rakende woordeboekgebruiksopleiding volg nie, alhoewel almal daarop aanspraak maak dat hulle die boek ontwerp het deur dit op die voorgeskrewe regeringskurrikulum te baseer. Voorstelle word gemaak oor hoe die huidige handboekontwerpe met die aanbevole hulpbronne woordeboekvaardighede kan inkorporeer. In die studie word die opvallende afwesigheid van opleiding in woordeboekgebruiksvaardighede in sekondêre handboeke blootgelê, en word soortgelyke studies in ander lande versoek sodat die (on)beskikbaarheid van woordeboekgebruiksopleiding in sekondêre skole ten volle begryp kan word. Hierdie studie behoort bruikbare inligting aan relevante regeringsowerhede, woordeboeksamestellers, handboekskrywers, en Engelse taalonderwysers en -navorsers te kan voorsien om die situasie te verbeter.

Sleutelwoorde: WOORDEBOEKGEBRUIKSOPLEIDING, TAIWAN, SEKONDÊRE SKOOL-EVT-HANDBOEK, INHOUDSANALISE

1. Introduction

The habit and the skills of the use of reference tools are instrumental to self-learning of a language. As with other autonomous learning abilities, such skills and habits should be trained as early as possible. However, dictionary use research overwhelmingly found that the training of dictionary use has been in neglect for decades (Herbst and Stein 1987, Nesi 2003, Li and Hai 2015). English teachers of the foreign language classroom, who have little prior training themselves, spare little time, with few ready-made materials, for students in this area. With a focus on the training resource on the secondary school level, this study, aims to examine the design of dictionary skills teaching in English textbooks in Taiwan for their effectiveness in following the guidelines set by the government, and how they could be improved for students to become more independent in language learning. It is hoped that this study could shed light on lexicographical pedagogy, and draw the attention of all stakeholders concerned, be they teachers, textbook compilers, and the government, for improvement.

2. Literature review

2.1 Research on dictionary use

Dictionary use research mostly focused on how tertiary students used the dictionary for English reading comprehension, with questionnaire survey and interviews being the major instruments, e.g., Chi 2003, Huang 2003, Thumb 2004, Li and Hai 2015, Campoy-Cubillo 2021. While Chan (2005) applied the think-aloud approach for her enquiry, some others used performance tasks to examine students' dictionary use strategies or needs for vocabulary learning (e.g., Chen 2012, Chan 2012, Esfandiari and Hezari 2019), or translation, e.g., Li 2003 and Law 2009, with the latter applying the think-aloud protocol, too. Tono (2003) discussed the methodological considerations on the research on dictionary use.

Research by user surveys (Law 2009, Chen 2010, Müller-Spitzer et al. 2012, Liu, Zheng and Chen 2019, Ma 2019) on other media of the dictionary, i.e. electronic, online, or mobile dictionaries found that these dictionaries are gaining acceptance among students, and have become the common reference tools to find or check the meaning of English words. These empirical studies established and affirmed the educational value of electronic dictionaries in L2 learning, and identified both positive features and contribution to L2 learning. However, in terms of research validity, Chi (2013: 175) commented that "when the reported data of these studies are to be referenced for decisions by a dictionary compiler, which may claim to meet the needs of millions…. [in Hartmann's (2001: 94) words], 'the results of various studies are of limited generalizability'."

In terms of subjects, although dictionary use skill is commonly regarded as elementary to basic learning, there are strangely far fewer studies on secondary school students vis-à-vis their dictionary use patterns (e.g., Wang and Wei 2009, Investigation Team 1999 targeting other professionals, too). In understanding user needs, Kwary (2018) stated that only a small handful of references have discussed the variables to be considered in determining dictionary user profiles. Other studies on this aspect included Chan 2011, Hamouda 2013, Lai and Chen 2015, while Mavrommatidou et al. (2019: 401-402) developed and validated the strategy inventory for electronic dictionary use with 32 items.

Overall, the findings of these studies and others (e.g., Chi 2011) revealed that learners in general are ignorant of what can be found in dictionaries, and of how to use them to meet their needs. One of the major reasons is the lack of dictionary skills training in secondary and tertiary education (cf., Chi 2003, Li 2003, Li and Hai 2015, Lopera 2019, Campoy-Cubillo 2021). One direct negative consequences of such lack is inadequate English vocabulary for study. Both McNeill's (2006) and Chi's (2011) research evidenced that Hong Kong university entrants generally fell far short of the 5000 words necessary to cope with university study in English, although they have studied under the local English language curriculum for twelve years in primary and secondary schools.

In the light of these findings, the call for dictionary skills training through integrating the learning topics in the language curriculum is almost unanimous among researchers and students (cf., Lan Li 1998, Investigation Team 1999, Miller 2008, Chi 2011), including doing regular dictionary usage exercises on authentic tasks (Herbst and Stein 1987, Wong 1996, Wingate 2004), and introducing to students the various types of dictionaries available in the market, in addition to bilingualized ones (Chi 2003). Similar conclusions are also drawn from studies in the U.K. (Barnes, Hunt and Powell 1999, Wingate 2002, 2004), and Japan (Mochizuki 2011). Official syllabuses for secondary schools in Hong Kong (Curriculum Development Council 1999) stipulated that using references such as the dictionary and thesaurus is essential to independent learning. Similar recommendations are found in the National Curriculum for England (Wingate 2004).

2.2 Dictionary skills training

In response to numerous calls for encompassing training of dictionary use in school and academic curricula (e.g., Campoy-Cubillo 2015, Lopera 2019, Al-Harbi 2024, Bothma and Fourie 2024), some workbooks and other materials aiming to improve dictionary users' reference skills have appeared (see Stark 1990 for an informative overview; Kipfer 1984, Chi 2003), and actual research into

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1880 (Article)

the effectiveness of training in dictionary reference skills have been carried out gradually in recent decades, e.g., Bae 2011, Cote González and Tejedor Martínez 2011, Liu 2014, Prćić 2020, Al-Harbi 2024, Bothma and Fourie 2024. However, these workbooks tend to be dictionary-specific and "emphasize the advantages of their attendant dictionaries to the almost total exclusion of any shortcomings" (Stark 1990: 4).

In the few studies targeting primary and secondary school EFL learners, Kipfer (1984) used a workbook accompanying the *Longman Dictionary of Contemporary English* to train American high-school students in dictionary use, and concluded that the instruction was effective. Lew and Galas (2008) measured the effectiveness of lexicographic training for primary-school-level Polish learners of English prior to and following a 12-session specially designed training programme. Pousi (2010) trained ninth grade EFL class dictionary use in Finland, and studied the effect on the dictionary use of her pupils. These studies predominantly demonstrated the positive results from trainees after the training, who became more effective in dictionary use and in self-learning, compared with the control group. However, on the teachers' end, the scenario is not very encouraging. Over the generations, much as their students, teachers have barely received any dictionary skills training themselves (cf., Herbst and Stein 1987, Oh 2006, Chi 2011). They may assess dictionary consultation as inferior to contextual guessing (Wingate 2002, 2004).

Another barrier to the learning of dictionary skills at school seems to be the limited amount of time spent on teaching them. The topic is commonly considered peripheral in foreign language teaching (Herbst and Stein 1987). Even if teachers could spare the time for such training, "up until the present moment, there is no framework or guidance to advise teachers how to teach dictionary use at junior secondary level. Such teaching at school, if it exists, has been incidental and sporadic, depending mostly on individual teacher's judgment" (Chi 2011: 80). While electronic dictionaries are becoming more popular among students, many teachers are either unfamiliar and/or skeptical about them (e.g., Taylor and Chan in Nesi 1999, Boonmoh 2010). Study results indicated that students would spontaneously look up a large number of unknown or unclear words with the device simple out of curiosity (Guillot and Kenning in Nesi 1999: 64). Chi (2020) established the need to introduce the use of dictionary training into formal EFL teaching. One suggestion to bridge the gap between the dictionary and EFL teachers is to provide training to English language teachers on "dictionary literacy".

2.3 Research on the textbook in dictionary skills training

The textbook works as a major determiner of schools' curricular, helps teachers organize their teaching, and is the main instructional material in class (Yen 2000). For learners, it may be the most important language input other than that from the teacher (Lee and Bathmaker 2007). It can be used to enhance learners' vocabulary building, reception, and production skills, and should thus be treated as an integral part of foreign language teaching (Herbst and Stein 1987; Law 2009). Normally, the training of dictionary use skills is incorporated into the textbook. Some dictionary publishers produce companion workbooks to their dictionaries. Stand-alone workbooks for dictionary skills are scanty. Stark's (1990) work remains a rare study on textbooks for dictionary use training. He proposed that the training should emphasize application to overcome linguistic problems rather than on the dictionary per se. Thornbury (2002) mentioned dictionary use training as one of the ways to teach vocabulary, and this topic is often included in course books.

While dictionary use training does not necessarily require a textbook, it is much easier and systematic to be carried out by teachers with a textbook designed according to the curriculum. It could consistently reach a large group of students at their elementary level. However, compared with the robust research on dictionary use and training shown above, study on dictionary skills training based on a textbook is scarcely explored. In July 2022, this researcher made a simple bibliometric search with keywords like 'textbook,' 'dictionary,' and their Chinese counterparts ('教科書', '詞典', '字典') at the following English or Chinese specialized journal websites: *Lexicographica* (https://www.degruyter.com/journal/key/lexi/html?lang=en), *Lexicography* (https://journal.equinoxpub.com/lexi/about), *Lexikos* (https://lexikos.journals.ac.za/pub), and the three associations' conference proceedings; two other prominent journals on lexicography: *International Journal of Lexicography* (https://academic.oup.com/ijl), and *Cishu yanjiu* (*Lexicographical Studies*; 中國期刊全文數據庫 (sris.com.tw)), as well as two major websites with academic bibliography: Google Scholar (Google Scholar) and WorldCat (WorldCat.org). Further searches were narrowed down to Taiwanese platforms. Airiti Library (Airiti Library) is a significant provider of digital academic bibliography, and *Curriculum & Instruction Quarterly* (see on Airiti Library) a major journal on the eponymous subject in Taiwan. Predictably, the searches delivered no results. This is a huge gap in dictionary use research which this study aims to fill.

Williams (1983) developed criteria for English language textbook evaluation: up-to-date methodology of L2 teaching, guidance for non-native speakers of English, needs of learners, relevance to socio-cultural environment in different aspects. Sheldon (1988: 242) proposed some other criteria. Tsai (1995) examined the varied aspects of an ESL/EFL textbook from a user's perspective. Tzong-Wei Li (1998) discussed the editing mechanism of textbook with particular considerations. Chou (2005) made a critical study of textbook research in Taiwan from 1979 to 2004 (cf. Tsai 2015). Chou (2005) pointed out that content analysis was the earliest and most popular method, while the quantitative approach the least common. Although not targeting the Taiwan context, Fan (2013) reviewed many studies on the textbook, and came to the same conclusion that their methodology was primarily content analysis, and also textbook comparison (see Meunier and Gouverneur 2007). Using both quantitative and qualitative con-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1880 (Article)

tent analysis of Taiwan's high-school English textbooks, Ke (2012) investigated the projected roles of English in Taiwan's high school English textbooks over the past 50 years. Chou and Cheng (2016) stated that the limitation of content analysis is that one cannot infer the intentions or the reception from the text content itself, as both depend on situational factors. To this, Shieh (2016) proposed an integrated approach of textbook research. Fang et al. (2015) targeted electronic textbooks for elementary and secondary schools.

Law (2013) examined the dictionary use instruction of some secondary Chinese language textbooks in Hong Kong, and concluded that although basically adhering to the guidelines of the education department of the government, the books nonetheless spare far from desirable space on it. This study is similar to Law's, yet on the English counterpart in a different geographical setting, so as to cover more ground in this subject.

2.4 EFL in Taiwan

Taiwan recognises written Chinese and its spoken form Mandarin Chinese as official language. Presently, Taiwanese students start to learn English as a Foreign Language (EFL) since grade three in primary school up to grade 12 in senior high school. The total study time is 10 years, although there are only one to two periods (45 minutes each) a week in grades three to six (Ministry of Education 2014: 15).

The decision to introduce English learning on the elementary school level in 2001, and the elimination of senior high and college entrance examinations effective in 2001 and 2002, respectively, are among the more important moves in the teaching of EFL in Taiwan. In addition, to reflect the features of communication-based teaching and to guide instruction material development and classroom practice, the Ministry of Education (MOE) has published new curricula for English teaching in both junior and senior high schools (Wang 2000). These new curricula have become a guiding concept for curriculum development and syllabus design, and sparked an array of language teaching innovations in EFL contexts in Taiwan (Wang and Savignon in Huang 2003).

3. Methodology

Notwithstanding the sets of assessment criteria for textbooks mentioned above (cf. Williams 1983, Sheldon 1988, Tsai 1995, Tzong-Wei Li 1998), this study will not adopt them for analysis, as detailed as they are, since our purpose is not thorough examination of the worth of a textbook, but a particular area of skills. In evaluating textbooks, Fan (2013: 771) analyzed that there are three types of variables: independent, intermediate, and dependent (Figure 1). This study will focus on the former two: the curriculum set by the government, which would affect the design of textbooks, and the textbooks themselves as in intermediate variable.

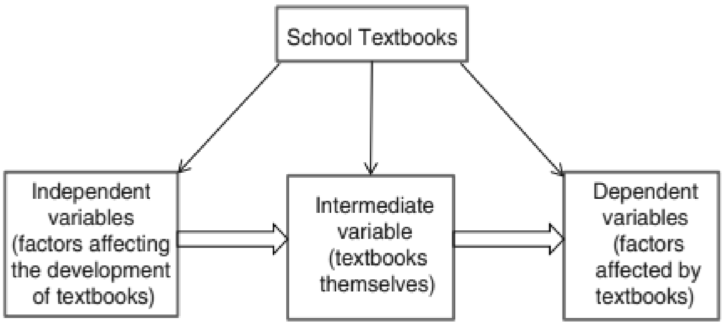

School Textbooks
Independent variables (factors affecting the development of textbooks)
Intermediate variable (textbooks themselves)
Dependent variables (factors affected by textbooks)

Figure 1: Textbooks viewed as an intermediate variable in the context of education

McDonough and Shaw's (in Meunier and Gouverneur 2007: 124) concepts of external and internal evaluation will be used in the study framework. What they two called "external" or "macro-evaluation," is what the books say about themselves, or what the author and/or publisher explicitly states as to the organisation and content of the book. The external evaluation can be achieved by analyzing the claims made on the front or back covers of the teacher's or student's book, in the introduction section, and finally in the table of contents of each textbook. The internal or micro-evaluation of a textbook consists in "analys[ing] the extent to which the […] factors in the external evaluation stage match up with the internal consistency and organization of the materials as stated by the author/publisher — for […] strong claims are often made for these materials" (ibid.: 125).

In a qualitative approach by document analysis, this study aims to find out:

(1) How effectively can the existing secondary English textbooks help learners develop the necessary dictionary skills based on the guidelines from the government; and
(2) How the existing textbooks could be improved to better meet learners' needs for dictionary skills training.

Document analysis is a systematic procedure for reviewing or evaluating documents, to elicit meaning, gain understanding, and develop empirical knowledge. It yields data that are organized into major themes, categories, and case examples specifically through content analysis (Bowen 2009: 27-28).

The evaluation criteria are based on the Curriculum Guidelines of the Ministry of Education (2018) (henceforth Guidelines), rendered in English from the original Chinese, and tabulated below in Table 1 for easy reference. Among the five curriculum objectives for the English Language (Ministry of Education 2018: 4),

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1880 (Article)

the pertinent one is: "to build up effective learning strategies for the English language, to strengthen self-learning capability, and to lay the foundation for lifelong learning".

Learning stage	Students' learning performance by nine learning and assessment factors
Third (grades 5-6)	7. Learning methods and strategies — can search for the pronunciation and meanings of words from dictionaries
Fourth (grades 7–9)	6. Learning interest and attitude (p. 14) — should actively make use of diverse search tools to understand the English language information they come across. 7. Learning methods and strategies (p. 14) — can make use of the dictionary to find suitable word meanings in context.
Fifth (grades 10–12)	6. Learning interest and attitude (p. 14) — can actively seek relevant English language learning resources from the Internet or other channels, and share with teachers and peers. 7. Learning methods and strategies (p. 15) — can actively use reference tools (e.g., dictionaries, encyclopaedias) or other online resources to understand the English content they come across

Table 1: Students' learning performance by grades in the English language curriculum as set by the MOE of Taiwan

The core attributes of the learning of the English language subject in Appendix 1 of the Guidelines (pp. 34-26) are the same as those listed in Table 1.

These guidelines are close to the six put forward by Mochizuki (2011). As well, in reference to Nesi's (2003) classification of dictionary reference skills, the guidelines coincide with a few of the skills mentioned: 1. Knowing what types of dictionary exist, and choosing which dictionary/ies to consult and/or buy; 2. Knowing what kinds of information are found in dictionaries and other types of references works; 3. Contextual guessing of the meaning of the look-up item; 4. Interpreting IPA and pronunciation information; and 5. Interpreting the definition or translation. But there is gross negligence of the skills for deciding the information for consultation, and locating the information wanted in an entry. The functions available on electronic or internet dictionaries are also not mentioned.

Conforming to all the guidelines does not necessarily mean that the textbook is best in design in dictionary use training. It only meets the basic curric-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1880 (Article)

ular expectation of the government in English language teaching. It is hoped that, with the findings, textbook publishers, teachers, and lexicographers could help improve the situation by providing better training methods and material for learners' needs.

The textbooks aimed at grades 7–12 of general secondary schools were selected since the student population was the highest among all the secondary school types: general, comprehensive, specialized, skills based (Ministry of Education 2022). The student number in junior secondary schools was 587,000 in 2021, while that in senior secondary schools was 586,000 in the same year, among which 280,000 came from general secondary schools on the junior and the senior levels respectively, about half of the total (Department of Statistics, Ministry of Education 2022). The following four textbook publishers had a market share of over 90% (Peng et al. 2015: 3). That means their books are representative, being used by a predominant proportion of secondary schools in Taiwan. These English textbooks are published by commercial publishers rather than by the Ministry of Education after 1999. At the end of 2001, only six sets of textbooks were used in the senior high schools in the greater Taipei area (Chen 2002). School textbooks are sold en masse in secondary schools, so no retail sale is available. The researcher could only catch what was sold in the second-hand market online, and thus was unable to obtain textbook copies for each grade and each publisher. Also, space disallows a thorough review of a series for all six years (junior high school 1 to senior high school 3) from the same publisher. But a glimpse of several items in a series published in recent years (2016 to 2021) revealed that the basic structure and design are consistent in the whole series. In the light of this, it is deemed that representation of a few grades from a diverse range of publishers could reflect the organisation and design rationale of secondary school textbooks currently in use in Taiwan. The samples below will be evaluated externally as per their claims in the prefaces, and internally in the light of the MOE guidelines. For each book, one unit will be sampled for its organisation.

Below is the list of sampled textbooks:

(1) Lin, Chia-fang (Ed.). 2019. *English Language for Junior High School (Teaching Materials)*. Vol. 1. Tainan: Han Lin Publishing Co. (Han Lin 1 hereafter)
(2) Lin, Chia-fang (Ed.). 2021. *English Language for Junior High School (Teaching Materials)*. Vol. 5. Tainan: Han Lin Publishing Co. (Han Lin 5 hereafter)
(3) Lin, I-ping, Chang, Hui-wen and Wu, I-chen (Eds.). 2011. *English Language for Junior High School*. Vol. 5. Tainan: Kang Huan Educational Publishing. (Kang Huan 5 hereafter)
(4) Che, Pei-chun (Ed.). 2017. *English Language for General Senior High School*. Vol. 2. Taipei: Sanmin Book Co. (Sanmin 2 hereafter)
(5) Che, Pei-chun (Ed.). 2016. *English Language for General Senior High School*. Vol. 3. Taipei: Sanmin Book Co. (Sanmin 3 hereafter)
(6) Chou, Chung-tien (Ed.). 2017. *English Language for General Senior High School*. Vol. 5. New Taipei City: Lungteng Education Co. (Lungteng 5 hereafter)

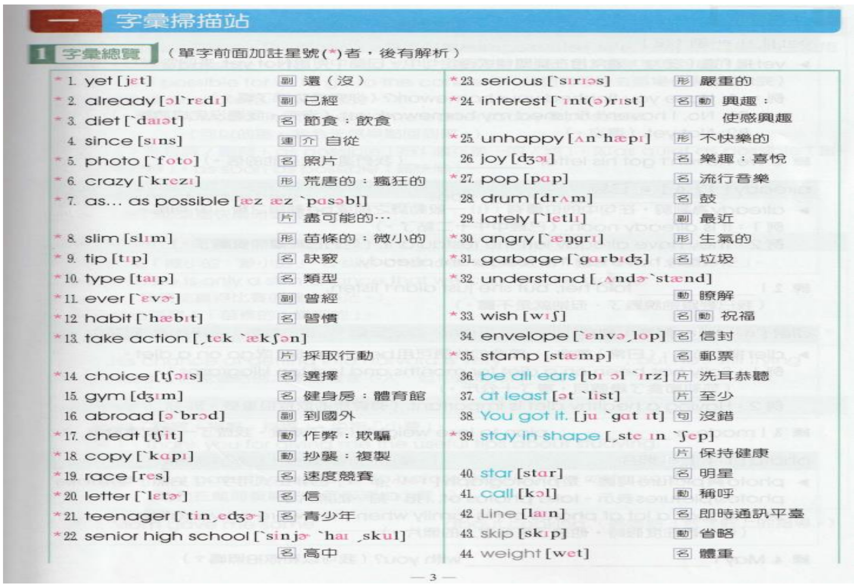
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1880 (Article)

(7) Chou, Chung-tien (Ed.). 2017. *English Language for General Senior High School*. Vol. 6. New Taipei City: Lungteng Education Co. (Lungteng 6 hereafter)

Items (1) to (3) are for junior high school, while (4) to (7) for senior. The external materials of the six samples are all written in Chinese, and rendered into English by the researcher himself.

4. Results and discussion

Han Lin 1 and Han Lin 5:
Before the main text, there is "How to use this book" to introduce the features of each unit: "the Essentials," "Vocabulary scanner," "Grammatical exploration region," "Text reading room," "the Essence City," "Test diagnosis from wide-angle lens," and "Test revision". There is no mention of the Guidelines, nor the learning performance stated therein. The "Vocabulary scanner" in Unit 1 ("Have you had breakfast yet?") of Han Lin 5 (pp. 3-4) comprises pronunciation symbols, grammatical labels in Chinese, and Chinese "equivalents" (Figure 2). Then comes the explanation of vocabulary (pp. 4-10), with pronunciation symbols, grammatical labels in Chinese, Chinese "equivalents," usage, example sentences, Chinese explanation, and practice sentences (Figure 3). There is a conversation analysis (p. 20) of the idioms used: I'm all ears (Figure 4), with Chinese explanation, example sentences, as well as Chinese translation.



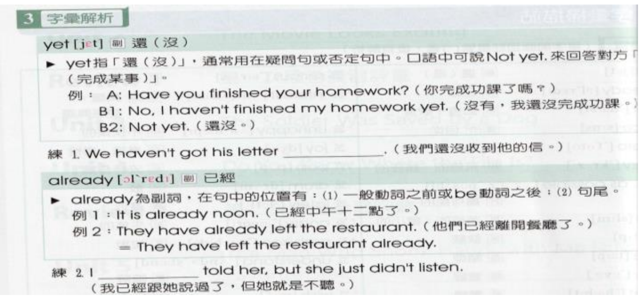
一 字彙掃描站

1 字彙總覽 （單字前面加註星號(*)者，後有解析）

*1. yet [jɛt]	副	還（沒）
*2. already [ɔl`rɛdɪ]	副	已經
*3. diet [`daɪət]	名	節食；飲食
4. since [sɪns]	連 介	自從
*5. photo [`foto]	名	照片
*6. crazy [`krezɪ]	形	荒唐的；瘋狂的
*7. as… as possible [æz æz `pɑsəbl̩]	片	盡可能的…
*8. slim [slɪm]	形	苗條的；微小的
*9. tip [tɪp]	名	訣竅
*10. type [taɪp]	名	類型
*11. ever [`ɛvɚ]	副	曾經
*12. habit [`hæbɪt]	名	習慣
*13. take action [ˏtek `ækʃən]	片	採取行動
*14. choice [tʃɔɪs]	名	選擇
15. gym [dʒɪm]	名	健身房；體育館
16. abroad [ə`brɔd]	副	到國外
*17. cheat [tʃit]	動	作弊；欺騙
*18. copy [`kɑpɪ]	動	抄襲；複製
*19. race [res]	名	速度競賽
*20. letter [`lɛtɚ]	名	信
*21. teenager [`tinˏedʒɚ]	名	青少年
*22. senior high school [`sinjɚ `haɪ ˏskul]	名	高中
*23. serious [`sɪrɪəs]	形	嚴重的
*24. interest [`ɪnt(ə)rɪst]	名 動	興趣；使感興趣
*25. unhappy [ʌn`hæpɪ]	形	不快樂的
26. joy [dʒɔɪ]	名	樂趣；喜悅
*27. pop [pɑp]	名	流行音樂
28. drum [drʌm]	名	鼓
29. lately [`letlɪ]	副	最近
*30. angry [`æŋgrɪ]	形	生氣的
*31. garbage [`gɑrbɪdʒ]	名	垃圾
*32. understand [ˏʌndɚ`stænd]	動	瞭解
*33. wish [wɪʃ]	名 動	祝福
34. envelope [`ɛnvəˏlop]	名	信封
*35. stamp [stæmp]	名	郵票
36. be all ears [bɪ ɔl `ɪrz]	片	洗耳恭聽
37. at least [ət `list]	片	至少
38. You got it. [ju `gɑt ɪt]	句	沒錯
*39. stay in shape [ˏste ɪn `ʃep]	片	保持健康
40. star [stɑr]	名	明星
41. call [kɔl]	動	稱呼
42. Line [laɪn]	名	即時通訊平臺
43. skip [skɪp]	動	省略
44. weight [wet]	名	體重

— 3 —

Figure 2: "Vocabulary scanner" in Unit 1 of Han Lin 5, p. 3



3 字彙解析

yet [jɛt] 副 還（沒）

► yet指「還（沒）」，通常用在疑問句或否定句中。口語中可說Not yet. 來回答對方「（完成某事）」。

例： A: Have you finished your homework?（你完成功課了嗎？）
B1: No, I haven't finished my homework yet.（沒有，我還沒完成功課。）
B2: Not yet.（還沒。）

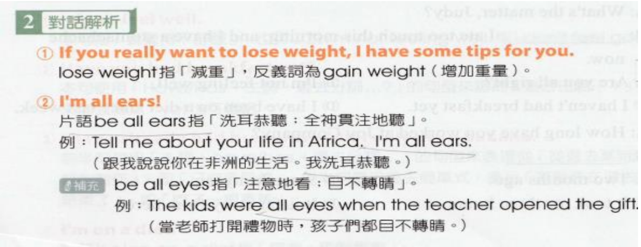
練 1. We haven't got his letter ________.（我們還沒收到他的信。）

already [ɔlˋrɛdɪ] 副 已經

► already為副詞，在句中的位置有：(1) 一般動詞之前或be動詞之後；(2) 句尾。
例 1：It is already noon.（已經中午十二點了。）
例 2：They have already left the restaurant.（他們已經離開餐廳了。）
= They have left the restaurant already.

練 2. I ________ told her, but she just didn't listen.
（我已經跟她說過了，但她就是不聽。）

Figure 3: Explanation of vocabulary, Unit 1, Han Lin 5, p. 4



2 對話解析

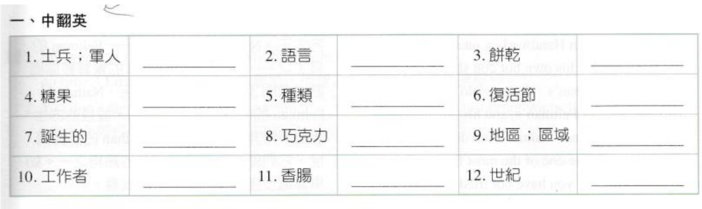
① If you really want to lose weight, I have some tips for you.
lose weight指「減重」，反義詞為gain weight（增加重量）。

② I'm all ears!
片語be all ears指「洗耳恭聽；全神貫注地聽」。
例：Tell me about your life in Africa. I'm all ears.
（跟我說說你在非洲的生活。我洗耳恭聽。）
補充 be all eyes指「注意地看；目不轉睛」。
例：The kids were all eyes when the teacher opened the gift.
（當老師打開禮物時，孩子們都目不轉睛。）

Figure 4: Conversation analysis, Unit 1, Han Lin 5, p. 20

<u>Kang Huan 5:</u>

Kang Huan 5 includes an "Editorial Intent," with an introduction of the organisation of each unit: learning objectives, vocabulary list, glossaries, try it, conversation and reading, grammar, essentials for examination, mock examination topics, review and reading level up. Yet it does not relate to the learning performance in the Guidelines. In Unit 1 ("Have you ever tried these dishes?"), there is a Vocabulary (p. 3) section, with pronunciation symbols, grammatical labels, and Chinese "equivalents". A later section called "vocabulary learning bar" (pp. 4-8) provides more explanation of the usage of those words, with example sentences, practice sentences, and Chinese translation. Near the end of the unit, there are two short exercises: translation of words from Chinese to English (p. 9) (Figure 5), and sentence translation from Chinese to English (p. 25) (Figure 6).

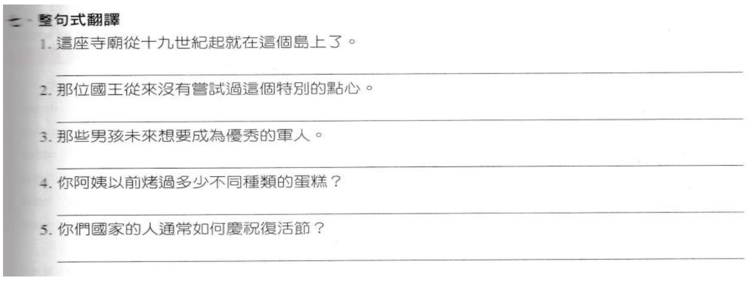


一、中翻英

1. 士兵；軍人	________	2. 語言	________	3. 餅乾	________
4. 糖果	________	5. 種類	________	6. 復活節	________
7. 誕生的	________	8. 巧克力	________	9. 地區；區域	________
10. 工作者	________	11. 香腸	________	12. 世紀	________

Figure 5: Word translation from Chinese to English, Unit 1, Kang Huan 5, p. 9

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1880 (Article)

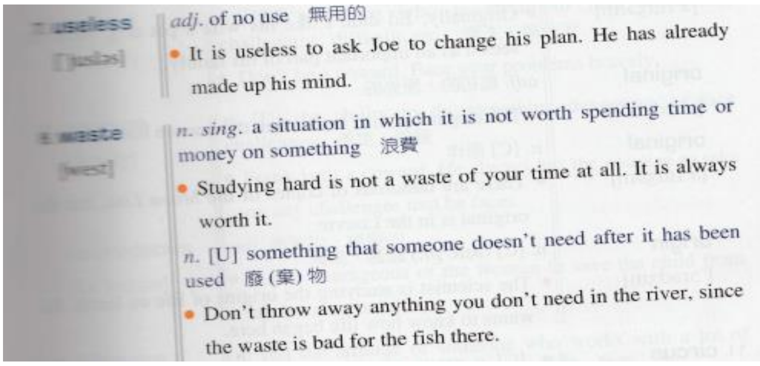


二、整句式翻譯

1. 這座寺廟從十九世紀起就在這個島上了。
2. 那位國王從來沒有嘗試過這個特別的點心。
3. 那些男孩未來想要成為優秀的軍人。
4. 你阿姨以前烤過多少不同種類的蛋糕？
5. 你們國家的人通常如何慶祝復活節？

Figure 6: Sentence translation from Chinese to English, Unit 1, Kang Huan 5, p. 25

<u>Sanmin 2 and Sanmin 3:</u>
The "Editorial Intent" in Sanmin2 says the book is compiled following the Guidelines (2018), but there is no mention of the students' learning performance. It lays out the structure of each unit: I. Reading (before you read, reading task, reading, after you read, words in use, patterns in use, beyond the text, writing hands-on); II. Listening and speaking (pp. i-iv). Unit One ("Animal imagery in language") includes KK pronunciation symbols (p. v), and abbreviations of grammatical symbols (p. vi). For "Words in Use: Words for production," difficult vocabulary found in the reading text is explained one by one. On pages 5-9, there are grammatical and pronunciation symbols, definitions in English, Chinese "equivalents," and example sentences (Figure 7).



7 useless [ˈjuslɪs]
adj. of no use 無用的
- It is useless to ask Joe to change his plan. He has already made up his mind.

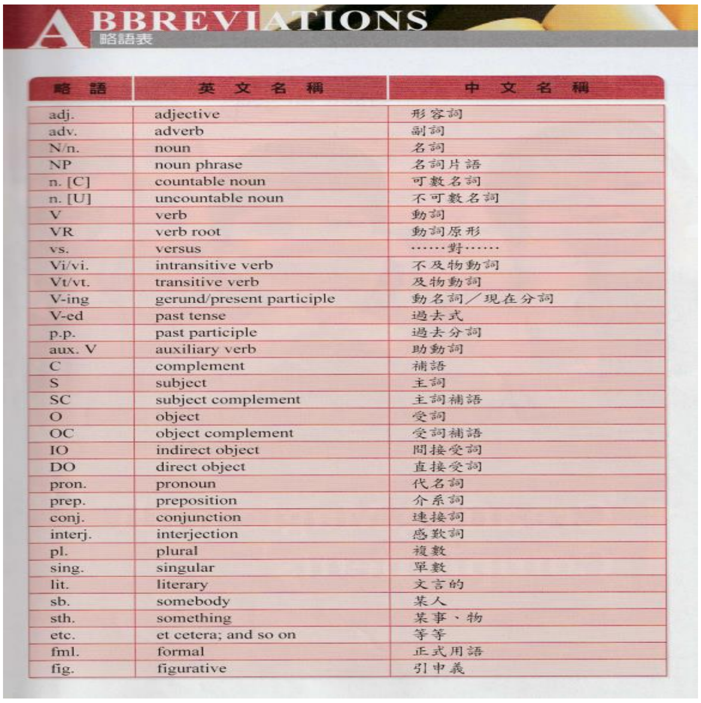
8 waste [west]
n. sing. a situation in which it is not worth spending time or money on something 浪費
- Studying hard is not a waste of your time at all. It is always worth it.

n. [U] something that someone doesn't need after it has been used 廢(棄)物
- Don't throw away anything you don't need in the river, since the waste is bad for the fish there.

Figure 7: Vocabulary explanation, Unit 1, Sanmin 2, p. 7

<u>Lungteng 5 and Lungteng 6:</u>
Lungteng 5 contains an "Editorial Intent" (pp. iv-v), which states that the organization design follows the Guidelines. The structure of each unit is like this: I. Reading (warm-up, reading selection, vocabulary, idioms and phrases, word

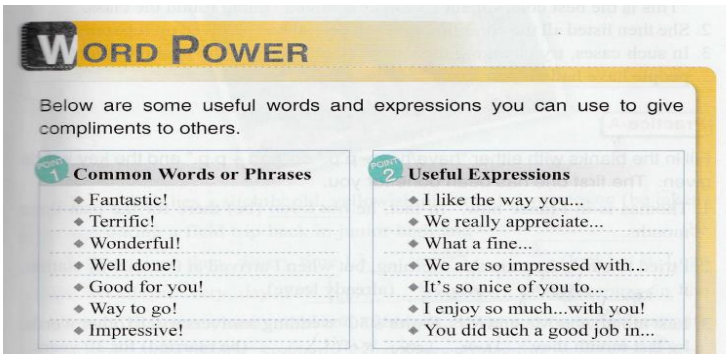
power, patterns in action, language in 1 & 2), guided conversations, on your own). In the main text, there is a "Pattern in action" (pp. vi-vii, ix), with grammatical symbols (Figure 8). In Unit One ("Creating magic with compliments"), there are "Difficult vocabulary" with pronunciation symbols (p. 3), vocabulary (pp. 7-12), idioms and phrases (p. 12). The "Word power" on page 13 shows synonyms (Figure 9).



ABBREVIATIONS
略語表

略 語	英 文 名 稱	中 文 名 稱
adj.	adjective	形容詞
adv.	adverb	副詞
N/n.	noun	名詞
NP	noun phrase	名詞片語
n. [C]	countable noun	可數名詞
n. [U]	uncountable noun	不可數名詞
V	verb	動詞
VR	verb root	動詞原形
vs.	versus	……對……
Vi/vi.	intransitive verb	不及物動詞
Vt/vt.	transitive verb	及物動詞
V-ing	gerund/present participle	動名詞／現在分詞
V-ed	past tense	過去式
p.p.	past participle	過去分詞
aux. V	auxiliary verb	助動詞
C	complement	補語
S	subject	主詞
SC	subject complement	主詞補語
O	object	受詞
OC	object complement	受詞補語
IO	indirect object	間接受詞
DO	direct object	直接受詞
pron.	pronoun	代名詞
prep.	preposition	介系詞
conj.	conjunction	連接詞
interj.	interjection	感歎詞
pl.	plural	複數
sing.	singular	單數
lit.	literary	文言的
sb.	somebody	某人
sth.	something	某事、物
etc.	et cetera; and so on	等等
fml.	formal	正式用語
fig.	figurative	引申義

Figure 8: Grammatical abbreviations, Lungteng 5, p. ix



WORD POWER

Below are some useful words and expressions you can use to give compliments to others.

POINT 1 Common Words or Phrases	POINT 2 Useful Expressions
◆ Fantastic!	◆ I like the way you...
◆ Terrific!	◆ We really appreciate...
◆ Wonderful!	◆ What a fine...
◆ Well done!	◆ We are so impressed with...
◆ Good for you!	◆ It's so nice of you to...
◆ Way to go!	◆ I enjoy so much...with you!
◆ Impressive!	◆ You did such a good job in...

Figure 9: Word power, Unit 1, Lungteng 5, p. 13

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1880 (Article)

In summary, most series highlight the adherence to the Guidelines in general organization. But in regard to dictionary use skills, none has complied. They do comprise something that can be found in a dictionary entry, e.g., pronunciation symbols, grammatical labels, example sentences, definitions, and Chinese "equivalents". Yet they fall short of encouraging users to search for the information themselves from the dictionary or to interpret the information from any dictionary entry. This is not desirable for independent learning.

The following table summarizes the sample series' (lack of) adherence to the Guidelines in assisting students to achieve learning performance by nine learning and assessment factors.

Guideline / Textbook series	Han Lin (vols. 1, 5)	Kang Huan (vol. 5)	Sanmin (vols. 2, 3)	Lungteng (vols. 5, 6)
Grades 7–9				
(1) Students should actively make use of diverse search tools to understand the English language information they come across.	X	X	N.A.	N.A.
(2) Students can make use of the dictionary to find suitable word meanings in context.	X	X	N.A.	N.A.
Grades 10–12				
(3) Students can actively seek relevant English language learning resources from the Internet or other channels, and share with teachers and peers.	N.A.	N.A.	X	X
(4) Students can actively use reference tools (e.g., dictionaries, encyclopaedia) or other online resources to understand the English content they come across.	N.A.	N.A.	X	X

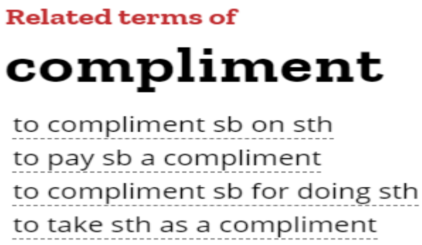
Table 2: An evaluation of the design of dictionary skills training in four series of secondary English textbooks mostly based on the Guidelines set out by the MOE. An "X" denotes unachievable

All textbooks are organised in units according to topics. Within these units, sections are devoted to the four (or five, if translating is included) skills, but also to grammar and vocabulary. Surprisingly, none of the surveyed series conforms to the Guidelines and devotes any space to fostering users' dictionary skills. One of the possible reasons is that the compilers take into consideration teachers' perspective more than the government's. They are used to drill-focused teaching practices in English language learning (Hsieh 2011). Relying on the textbook for full information enables both teachers and students a sense of control. Resorting to the dictionary and other online resources for answers means opening to a wide range of possibilities, without definite answers. The other possible reason may be the consistent neglect of such training in schools, no matter in Chinese language education (see Ta'ai 2006), or English one. Another barrier seems to be the limited subject lesson time. The topic is commonly considered peripheral in foreign language teaching (Herbst and Stein 1987). Even if teachers could spare the time for such training, they are not provided with many resources (cf. Chi 2011: 80). Other possible reasons may be found in Wingate (2002, 2004) and Liu (2014).

5. Recommendations and conclusion

The following recommendations are made based on the original textbook series for the learning performance in accordance with the Guidelines.

Almost all series contain pronunciation symbols for students to pronounce words accurately in accordance to the learning performance set in the Guidelines for grades 5–6. One easy (and sometimes free) way for students to imitate and practise is to go to online dictionaries and/or dictionary apps to listen to the demonstration. For other information of a word, the compilers can choose not to provide it, but to direct users to check from a certain online dictionary, e.g., the *Collins English–Chinese Dictionary* (https://www.collinsdictionary.com/dictionary/english-chinese). For example, for "compliment" in Lungteng 5, teachers could further encourage students to learn about the usage of the word by directing them to an online dictionary, such as *Collins English Dictionary* (see figure 10). Both dictionaries can be used by secondary school students of intermediate level of English.



Related terms of

compliment

to compliment sb on sth
to pay sb a compliment
to compliment sb for doing sth
to take sth as a compliment

Figure 10: The usage of the word "compliment" from the *Collins English Dictionary*

Other useful online and free reference tools include:

- *Chinese–English Dictionary* (https://cdict.net/)
- *Line Dictionary* (https://dict.naver.com/linedict/#/cnen/home)
- *Reverso Context (*https://context.reverso.net/translation/)
- *WordReference.com* (http://www.wordreference.com)
- *Youdao Dictionary* (http://dict.youdao.com/?keyfrom=cidian)

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1880 (Article)

Compared to the bilingual glossary with limited information provided in their textbooks, these dictionaries are certainly more complicated. That is exactly where teachers' instruction is needed. It is advised that teachers choose just one to two of them for demonstration. After this, it is hoped that learners can learn to be independent dictionary users, with occasional consultation with teachers. Thus, they can become more autonomous in lexical learning.

Instead of providing all the difficult words on the vocabulary list in a unit, the compiler can leave a few words for students to check up with the dictionary for meanings and other information. Students can also decide on which definitions to fit the context in the reading passage. To arouse interest, teachers could ask students for synonyms, e.g., "all ears" from Han Lin5 (cf. all ears | Synonyms and analogies for all ears in English | Reverso Dictionary, Thornbury 2002). As for translation exercises, given the popularity of online translation software even on dictionary websites, it may be impossible for teachers to forbid their use. Instead, teachers could use the software generated translation as the starting point for teaching how to edit the target text semantically and linguistically. This is where teacher input is necessary. In reference to Nesi's (2003) classification of dictionary reference skills, the textbooks could introduce bilingualized learner's dictionaries by major dictionary publishers like Oxford, Longman, Collins, which are popular among intermediate and advanced learners.

The suggestions above demonstrate how, with a little more attention to and resources for dictionary use training design, learners can practise more of their skills in relation to the unit theme with genuine tasks in a textbook. It does not occupy much space, but with recurrent practice in the whole textbook series, the skills can be consolidated, and the skills more sophisticated. While publishers may not be quick in revising their textbooks for this purpose, teachers can design their own worksheets based on the unit theme of the textbook. At the same time, students can still use bilingual dictionaries to assist in translation between English and Chinese during a reading, writing, or even speaking task. Teachers need to inform students of the potential benefits of an English-bilingualised dictionary, which is a hybrid of an English monolingual and an English–Chinese dictionary. Third, and probably most importantly, teachers should tell students what other options than dictionary use are available when encountering unknown words in reading English, or when unsure how to express an idea in English writing, or even speaking (Huang 2003).

None of the textbook series followed the Guidelines regarding dictionary use skills. Without ready-made teaching material, it is no wonder that teachers neglect them, and students are untrained. The results of the serious lack of dictionary use training in schools coincide with those found in mainland China (Investigation Team 1999), in Taiwan for the Chinese Language subject (Ta'ai 2006), and in Hong Kong (Law 2013). There is little room for dictionary use training in the secondary curriculum. When only the basic skills are taught, if at all, throughout students' secondary education, the more advanced skills would only be taken up by trial and error by students themselves. This means that plenty of them will probably not ever get a good grip on dictionary skills, which is far from

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1880 (Article)

desirable for language learning, learning independence, and lifelong learning. It is hoped that this study can probe the parties concerned, be it the government, textbook publishers, teachers, and pedagogical lexicographers, to consider how to make better use of the textbook for such training in the future. In future studies, the reasons for textbook publishers not to include such training in their books are worth investigation.

To ensure that schools assign adequate time on dictionary use training, Herbst and Stein (1987) proposed that dictionary skills be tested in school to force the student to acquire the full range of dictionary skills, and also oblige the teacher to teach them. Law (2009) put forward the notion that dictionary use competence be considered one of the many foundational learning skills in a learner's life since secondary school up to university, and a threshold be set each year to ascertain that learners have reached them, in a way like English, Mathematics, and Computer Literacy.

With this reference study, teachers could devise suitable dictionary use activities to be conducted in the classroom. Based on this rare study on the design of dictionary use training in secondary textbooks, researchers of lexicography could consider how to further develop lexicographical pedagogy. One direction for further study is to measure the effectiveness of such training. Thumb's study (2004) focused on dictionary skills training for reception for university students, while Bishop's (2001) was on writing, both of which reported positive results, with a methodology for measurement. Another direction is a comprehensive pedagogy for training from junior secondary school to university. Chi (2011) formulated a framework for dictionary use teaching in local junior secondary schools: (1) pre-test; (2) treatment: selected dictionary use items be integrated into the existing school English syllabus and taught explicitly; (3) post-test: collecting feedback for evaluation and verification, with an implementation process over 12 months. More detailed pedagogy could be developed to help teachers, who were mostly untrained in dictionary use before, to gain more ideas about how training could be conducted, and collated with other language skills. Béjoint (1989), Bishop (2001), Nesi (2003), and Law (2009), Gavriilidou (2013) have proposed what could be covered in training in school or university. And lastly, similar studies could be conducted in other countries with their (English) language textbooks pertinent to dictionary skills training. Although EFL teachers seem to be aware of students' dictionary perceptions and their dictionary strategy use, their own beliefs about dictionaries may differ from students', and students' actual dictionary use behaviors may run counter to teachers' expectations (Huang 2003). Future studies could explore the differences between teachers' and students' expectations in dictionary use, and how this would affect the practice of dictionary use training.

6. References

Al-Harbi, Hanan Habis. 2024. Investigating EFL Students' and Instructors' Perceptions of Dictionary Usage in Writing Assessment. *English Language Teaching* 17(2): 67-80.

Bae, Susanna. 2011. Teacher-training in Dictionary Use: Voices from Korean Teachers of English. Akasu, K. and S. Uchida (Eds.). 2011. *ASIALEX 2011 Proceedings. Lexicography: Theoretical and Practical Perspectives. Papers Submitted to the 7th ASIALEX Biennial International Conference:* 46-55. Kyoto: The Asian Association for Lexicography.

Barnes, Ann, Marilyn Hunt and Bob Powell. 1999. Dictionary Use in the Teaching and Examining of MFLs at GCSE. *The Language Learning Journal* 19: 19-27.

Béjoint, H. 1989. The Teaching of Dictionary Use: Present State and Future Tasks. Hausmann, Franz Josef, O. Reichmann, Herbert Ernst Wiegand et al. (Eds.). 1989. *Wörterbücher/Dictionaries/Dictionnaires*. Vol. 1: 208-215. Berlin/New York: Walter de Gruyter.

Bishop, Graham. 2001. Using Quality and Accuracy Ratings to Quantify the Value Added of a Dictionary Skills Training Course. *The Language Learning Journal* 24: 62-69.

Boonmoh, Atipat. 2010. Teachers' Use and Knowledge of Electronic Dictionaries. *ABAC Journal* 30(3): 56-74.

Bothma, Theo J.D. and Ina Fourie. 2024. Contextualised Dictionary Literacy, Information Literacy and Information Behaviour in the e-Environment. *Library Management*. https://doi.org/10.1108/LM-08-2023-0082

Bowen, Glenn A. 2009. Document Analysis as a Qualitative Research Method. *Qualitative Research Journal* 9(2): 27-40.

Campoy-Cubillo, Mari Carmen. 2015. Assessing Dictionary Skills. *Lexicography. Journal of ASIALEX* 2: 119-141.

Campoy-Cubillo, Mari Carmen. 2021. Fostering Learners' Online Dictionary Skills through Active Dictionary Rubrics. *Lexikos* 31: 487-510.

Chan, Alice Yin Wa. 2005. Tactics Employed and Problems Encountered by University English Majors in Hong Kong in Using a Dictionary. *Applied Language Learning* 15(1–2): 1-27.

Chan, Alice Yin Wa. 2011. Bilingualised or Monolingual Dictionaries? Preferences and Practices of Advanced ESL Learners in Hong Kong. *Language, Culture and Curriculum* 24(1): 1-21.

Chan, Alice Yin Wa. 2012. The Use of a Monolingual Dictionary for Meaning Determination by Advanced Cantonese ESL Learners in Hong Kong. *Applied Linguistics* 33(2): 115-140.

Chen, Chen-Ting. 2002. *Textbook Selection for Senior High School Students in Greater Taipei Area*. Unpublished Master's thesis, National Taiwan Normal University, Taipei.

Chen, Yuzhen. 2010. Dictionary Use and EFL Learning. A Contrastive Study of Pocket Electronic Dictionaries and Paper Dictionaries. *International Journal of Lexicography* 23(3): 275-306.

Chen, Yuzhen. 2012. Dictionary Use and Vocabulary Learning in the Context of Reading. *International Journal of Lexicography* 25(2): 216-247.

Chi, Amy. 2003. Teaching Dictionary Skills in the Classroom. Hartmann, R.R.K. (Ed.). 2003. *Lexicography — Critical Concepts. Volume 1: Dictionaries, Compilers, Critics and Users*: 355-369. London/ New York: Routledge.

Chi, Amy. 2011. When Dictionaries Support Vocabulary Learning, Where to Begin? Akasu, K. and S. Uchida (Eds.). 2011. *ASIALEX 2011 Proceedings. Lexicography: Theoretical and Practical Perspectives. Papers Submitted to the 7th ASIALEX Biennial International Conference:* 76-85. Kyoto: The Asian Association for Lexicography.

Chi, Amy. 2013. Researching Pedagogical Lexicography. Jackson, H. (Ed.). 2013. *The Bloomsbury Companion to Lexicography:* 165-187. London: Bloomsbury Academic.

Chi, Amy. 2020. Reconstructing the Lexicographical Triangle through Teaching Dictionary Literacy to Teachers of English. *Lexicography. Journal of ASIALEX* 7(1–2): 79-95.

Chou, Pei-I. 2005. A Critical Study of Textbook Research in Taiwan 1979–2004. *Curriculum & Instruction Quarterly* 8(4): 91-116.

Chou, Pei-I and M.-C. Cheng. 2016. Textbook Research Methodology: Towards a Pluralistic Paradigm. C*urriculum & Instruction Quarterly* 19(3): 1-26.

Cote González, Margarita and Cristina Tejedor Martínez. 2011. The Effect of Dictionary Training in the Teaching of English as a Foreign Language. *Revista Alicantina de Estudios Ingleses* 24: 31-52.

Curriculum Development Council. 1999. *Syllabuses for Secondary Schools — English Language, Secondary 1–5*. Hong Kong: Education Department, Government of HKSAR.

Department of Statistics, Ministry of Education. 2022. *Yiyi xuenian xuexiao jiben gaikuang tongji jiguo tiyao fenxi* [Summary Analysis of Statistical Results of Schools in 2021]. Ministry of Education. <4D6963726F736F667420576F7264202D2030315F313130BEC7AED5B0F2A5BBB7A7AA70B4A 3AD6EA4C0AA522E646F63> (moe.gov.tw)

Esfandiari, R. and M. Hezari. 2019. The Comparative Effect of Dictionary Use, Etymological Analysis, and Glossing on Iranian ESP Learners' Vocabulary Learning. *Taiwan International ESP Journal* 10(1): 45-63.

Fan, Lianghuo. 2013. Textbook Research as Scientific Research: Towards a Common Ground on Issues and Methods of Research on Mathematics Textbooks. *ZDM Mathematics Education* 45(5): 765-777.

Fang, Chi-Hua, Shing-Hua Yeh, Yu-Yang Liu and Shin-Rou Huang. 2015. Research on Trends and Requirements of e-Textbook Industry Developed for Elementary and Secondary Schools. *Curriculum & Instruction Quarterly* 18(3): 157-182.

Gavriilidou, Z. 2013. Development and Validation of the Strategy Inventory for Dictionary Use (S.I.D.U.). *International Journal of Lexicography* 26(2): 135-153.

Hamouda, Arafat. 2013. A Study of Dictionary Use by Saudi EFL Students at Qassim University. *Study in English Language Teaching* 1(1): 227-257.

Herbst, Thomas and Gabrielle Stein. 1987. Dictionary-Using Skills: A Plea for a New Orientation in Language Teaching. Cowie, A. (Ed.). 1985. *The Dictionary and the Language Learner: Papers from the EURALEX Seminar at the University of Leeds, 1–3 April 1985*: 115-127. Tübingen: Max Niemeyer.

Hsieh, Ming Fang. 2011. Learning English as a Foreign Language in Taiwan: Students' Experiences and Beyond. *Language Awareness* 20(3): 255-270.

Huang, Da-Fu. 2003. *Taiwanese University English Majors' Beliefs about English Dictionaries and Their Dictionary Strategy Use*. Unpublished Doctoral Dissertation, The University of Texas at Austin.

Investigation Team. 1999. Yuwen cishu shiyong qingkuang diaocha baokao [Investigative Survey Report on Language Dictionary Use]. *Cishu yanjiu* [Lexicographical Studies] 5: 140-148.

Ke, I-Chung. 2012. From EFL to English as an International and Scientific Language: Analysing Taiwan's High-School English Textbooks in the Period 1952–2009. *Language, Culture and Curriculum* 25(2): 173-187.
DOI: 10.1080/07908318.2012.683530

Kipfer, Barbara Ann. 1984. *Workbook on Lexicography — A Course for Dictionary Uses with a Glossary of English Lexicographical Terms.* Exeter: University of Exeter.

Kwary, Deny A. 2018. The Variables for Drawing up the Profile of Dictionary Users. *Lexicography. Journal of ASIALEX* 4(2): 105-118.

Lai, Shu-Li and Chen, Hao-Jan Howard. 2015. Dictionaries vs Concordancers: Actual Practice of the Two Different Tools in EFL Writing. *Computer Assisted Language Learning* 28(4): 341-363.

Law, Wai-on. 2009. *Translation Students' Use of Dictionaries: A Hong Kong Case Study for Chinese to English Translation.* Unpublished Doctoral thesis, The University of Durham.

Law, Wai-on. 2013. Xianggang Chuzhong Zhongwen Jiaokeshu Gongjushu Jiaoxue Sheji Chutang [A Preliminary Investigation of the Dictionary Teaching Design in Junior Secondary School Chinese Language Textbooks in Hong Kong]. *Current Research in Chinese Linguistics* 92(1): 37-43.

Lee, Rachel N.F. and Ann-Marie Bathmaker. 2007. The Use of English Textbooks for Teaching English to "Vocational" Students in Singapore Secondary Schools: A Survey of Teachers' Beliefs. *RELC Journal* 38(3): 350-374.

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, Barcelona, 15–19 July 2008:* 1273-1285. Barcelona: Universitat Pompeu Fabra.

Li, Defeng. 2003. Compilation of English–Chinese Dictionaries: The User's Perspective. *Journal of Translation Studies* 8: 91-115.

Li, Lan. 1998. Dictionaries and Their Users at Chinese Universities: With Special Reference to ESP Learners. McArthur, T. and I. Kernerman (Eds.). 1998. *Lexicography in Asia: Selected Papers from the Dictionaries in Asia Conference, Hong Kong University of Science and Technology, 1997 and Other Papers*: 61-79. Tel Aviv: Password Publishers.

Li, Tzong-Wei. 1998. The Editing Mechanism of Textbook. *Curriculum & Instruction Quarterly* 1(1): 41-55.

Li, Lingling and Hai Xu. 2015. Using an Online Dictionary for Identifying the Meanings of Verb Phrases by Chinese EFL Learners. *Lexikos* 25: 191-209.

Liu, Lin. 2014. The Integration of Dictionary Use Strategy Training into Basic English Class. *Theory and Practice in Language Studies* 4(10): 2138-2143.

Liu, Xiqin, Dongping Zheng and Yushuai Chen. 2019. Latent Classes of Smartphone Dictionary Users among Chinese EFL Learners: A Mixed-Method Inquiry into Motivation for Mobile Assisted Language Learning. *International Journal of Lexicography* 32(1): 68-91.

Lopera, Sergio A. 2019. Effects of Reading Strategy and Dictionary Instruction in an Undergraduate Foreign Language Reading Comprehension Group. *Folios* 50: 127-138. doi: 10.17227/Folios.50-10226.

Ma, Qing. 2019. University L2 learners' Voices and Experience in Making Use of Dictionary Apps in Mobile Assisted Language Learning (MALL). *International Journal of Computer-Assisted Language Learning and Teaching* 9(4): 18-36.

Mavrommatidou, Stavroula, Zoe Gavriilidou and Angelos Markos. 2019. Development and Validation of the Strategy Inventory for Electronic Dictionary Use (S.I.E.D.U.). *International Journal of Lexicography* 32(4): 393-410.

McNeill, A. 2006. *English Vocabulary in the Secondary Classroom.* https://www.edb.gov.hk/en/curriculum-development/kla/eng-edu/references-resources/resource%20sec_vocab.html

Meunier, Fanny and Céline Gouverneur. 2007. The Treatment of Phraseology in ELT Textbooks. Hidalgo, E., L. Quereda and J. Santana (Eds.). 2007. *Corpora in the Foreign Language Classroom: Selected Papers from the Sixth International Conference on Teaching and Language Corpora (TaLC 6), University of Granada, Spain, 4–7 July 2004:* 119-139. Amsterdam/New York: Brill.

Miller, Julia. 2008. Teachers and Dictionaries in Australia: Is There a Need to Train the Trainers? *TESOL in Context* 17(2): 11-19.

Ministry of Education. 2022. *Education in Taiwan*. Ministry of Education, Republic of China. Education in Taiwan (2022/2023) (moe.gov.tw)

Ministry of Education, Republic of China. 2014. *Curriculum Guidelines of 12-year Basic Education General Guidelines*.
Curriculum Guidelines of 12-year Basic Education-NAER | research. action. impact

Ministry of Education, Republic of China. 2018. *Curriculum Guidelines of 12-year Basic Education for Elementary, Junior High Schools and General Senior High Schools Language Arts — English* [in Chinese].
104.05.15會議議程版 本資料僅供此次會議參閱，會後請交由現場人員回收，謝謝！(naer.edu.tw)

Mochizuki, Masamichi. 2011. Guidelines on Dictionary Use Instruction. Akasu, K. and S. Uchida (Eds.). 2011. *ASIALEX 2011 Proceedings. Lexicography: Theoretical and Practical Perspectives. Papers Submitted to the 7th ASIALEX Biennial International Conference:* 368-377. Kyoto: The Asian Association for Lexicography.

Müller-Spitzer, Carolin, Alexander Koplenig and Antje Töpel. 2012. Online Dictionary Use: Key Findings from an Empirical Research Project. Granger, Sylviane and Magali Paquot (Eds.). 2012. *Electronic Lexicography*. Online Edition. Oxford: Oxford Academic.
https://doi.org/10.1093/acprof:oso/9780199654864.003.0020

Nesi, Hilary. 1999. A User's Guide to Electronic Dictionaries for Language Learners. *International Journal of Lexicography* 12(1): 55-66.

Nesi, Hilary. 2003. The Specification of Dictionary Reference Skills in Higher Education. Hartmann, R.R.K. (Ed.). 2003. *Lexicography — Critical Concepts. Volume 1: Dictionaries, Compilers, Critics and Users*: 370-393. London/New York: Routledge.

Oh, H.J. 2006. Korean English Teachers' Use of English Dictionaries and Instruction on their Use. *Foreign Language Education* 13(3): 419-443.

Peng, Chih-ling, Hsin-jou Huang and Shun-te Lan. 2015. Cong duoyuan dao guazhan: jiaokeshu shichang de tiaozhan yu yinying [From Diversity to Oligarchy: Challenges and Adaptations in the Textbook Market] *Taiwan Educational Review Monthly* 4(8): 1-5.
01主009. 彭致翎、黃欣柔、藍順德_從多元到寡占：教科書市場的挑戰與因應.pdf (ater.org.tw)

Pousi, A. 2010. *Training in Dictionary Use: A Teaching Intervention in a 9th Grade EFL Classroom in Finland.* Bachelor's thesis, University of Jyväskylä, Finland.

Prćić, T. 2020. Teaching Lexicography as a University Course: Theoretical, Practical and Critical Considerations. *Lexikos* 30: 293-320.

Sheldon, Leslie E. 1988. Evaluating ELT Textbooks and Materials. *ELT Journal* 42(4): 237-246.

Shieh, Jin-Jy. 2016. The Integrated Approach of Textbook Research: Mixed Methods Research. *Curriculum & Instruction Quarterly* 19(3): 55-83.

Stark, M.P. 1990. *Dictionary Workbooks — A Critical Evaluation of Dictionary Workbooks for the Foreign Language Learner.* Exeter: University of Exeter.

Ta'ai, Mei-hui. 2006. *Taiwan zhongxue guowen jiaoxue yanjiu* [Studies on the Teaching and Learning of the Chinese Language in Secondary Schools in Taiwan]. Guangzhou: Guangdong Education Publications.

Thornbury, Scott. 2002. *How to Teach Vocabulary*. Essex: Pearson.

Thumb, J. 2004. *Dictionary Look-up Strategies and the Bilingualised Learner's Dictionary — A Think-Aloud Study.* Tübingen: Max Niemeyer.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1880 (Article)

Tono, Yukio. 2003. Research on Dictionary Use: Methodological Considerations. Hartmann, R.R.K. (Ed.). 2003. *Lexicography — Critical Concepts. Volume 1: Dictionaries, Compilers, Critics and Users*: 394-412. London/New York: Routledge.

Tsai, K.-J. 2015. Profiling the Collocation Use in ELT Textbooks and Learner Writing. *Language Teaching Research* 19(6): 723-740.

Tsai, Y.-H. 1995. Textbook Selection for the EFL Classroom: The Students' Perspective. *Hwa Kang Journal of TEFL* (1): 1-30.

Wang, A. and X. Wei. 2009. Zhi qi suo xv, yonghu shou yi [An Understanding of User Needs Benefits Users]. *Cishu yanjiu* [Lexicographical Studies] 6: 117-125.

Wang, Chaochang. 2000. *A Sociolinguistic Profile of English in Taiwan: Social Context and Learner Needs*. Doctoral dissertation, The Pennsylvania State University.

Williams, David. 1983. Developing Criteria for Textbook Evaluation. *ELT Journal* 37(3): 251-255.

Wingate, Ursula. 2002. The Effectiveness of Different Learner Dictionaries. Tübingen: Max Niemeyer.

Wingate, Ursula. 2004. Dictionary Use — The Need to Teach Strategies. *The Language Learning Journal* 29*:* 5-11.

Wong, W.W.L. 1996. *Does the Use of the Dictionary Help Students in their Reading Comprehension and Cloze Exercises?* Unpublished Master's Dissertation, Hong Kong Polytechnic University Hung Hom, Kowloon.

Yen, Yuh-Yun. 2000. *Identity Issues in EFL and ESL Textbooks: A Sociocultural Perspective.* Unpublished Doctoral Dissertation. Ohio State University, Columbus.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1882 (Article)

How Can We Raise Strategic Dictionary Use in the Classroom: The Effect of a Dictionary Awareness Program on Dictionary Use Strategies

Zoe Gavriilidou, *Democritus University of Thrace, Greece*
(zoegab@otenet.gr) (https://orcid.org/0000-0002-5975-6852)

Angelos Markos, *Democritus University of Thrace, Greece*
(amarkos@eled.duth.gr) (https://orcid.org/0000-0002-4204-3573)
and
Evanthia Konstantinidou, *Democritus University of Thrace, Greece*
(evakons@helit.duth.gr) (https://orcid.org/0009-0001-3180-9916)

Abstract: This study investigates the impact of an explicit and integrated dictionary awareness program on primary school pupils' dictionary use strategies. The survey involved a total of 150 participants, aged 10–12 years old, from mainstream and intercultural schools. Data was collected before and after the implementation of the program using the Strategy Inventory for Dictionary Use (SIDU), a reliable and validated self-report tool that accurately profiles paper dictionary users' reported use in real-life contexts (Gavriilidou 2013). The dictionary awareness program consisted of targeted activities and was implemented to a group of 75 students, including 50 from mainstream schools and 25 from an intercultural school. The findings suggest that there is a lack of dictionary culture among students attending Greek schools, as evidenced by the moderate strategic use of dictionaries and the incomplete integration of dictionaries as reference tools in the educational process. Additionally, the comparison of the percentage of each strategy category before and after the implementation of the program showed a significant effect of the program on all categories of Dictionary Use Strategies (DUS) employed by the experimental group. This study contributes to the discussion of the "teachability" of dictionary use strategies by highlighting the effectiveness of dictionary awareness programs in promoting a dictionary culture.

Keywords: DICTIONARY USE STRATEGIES, DICTIONARY AWARENESS PROGRAM, EXPLICIT AND INTEGRATED STRATEGY INSTRUCTION, DICTIONARY CULTURE, CALLA, STRATEGY BASED INSTRUCTION, LOOK UP STRATEGIES, LEMMATISATION STRATEGIES

Opsomming: Hoe strategiese woordeboekgebruik in die klaskamer verhoog kan word: Die uitwerking van 'n woordeboekbewusmakingsprogram op woordeboekgebruikstrategieë. Hierdie studie ondersoek die impak wat 'n eksplisiete en geïntegreerde woordeboekbewusmakingsprogram op primêreskoolleerders se woordeboekgebruikstrategieë het. Die opname het altesaam 150 deelnemers, 10–12 jaar oud, van hoofstroom- en interkulturele skole betrek. Data is versamel voor en ná implementering van die program wat van die Strategie-inventaris vir Woordeboekgebruik (SIWG) gebruik gemaak het. Dit is 'n betroubare en bewese selfrapporteringshulpmiddel wat gebruikers van papierwoordeboeke se gerapporteerde gebruik binne die konteks van die werklike lewe akkuraat profilieer (Gavriilidou 2013). Die woordeboekbewusmakingsprogram het uit doelmatige aktiwiteite bestaan en is deur 'n groep van 75 studente, insluitende 50 uit hoofstroomskole en 25 uit 'n interkulturele skool, geïmplementeer. Die bevindings dui daarop dat daar onder studente wat Griekse skole bywoon 'n gebrek aan woordeboekkultuur is soos duidelik blyk uit die matige strategiese gebruik van woordeboeke en die onvolledige integrasie van woordeboeke as verwysingshulpmiddels in die opvoedkundige proses. Daarbenewens het die vergelyking van die persentasie van elke strategiekategorie voor en ná die implementering van die program aangetoon dat die program 'n beduidende uitwerking op alle kategorieë van Woordeboekgebruikstrategieë (WGS) wat deur die eksperimentele groep geïmplementeer is, gehad het. Hierdie studie lewer 'n bydrae tot die bespreking van die leerbaarheid van woordeboekgebruikstrategieë deur die effektiwiteit van woordeboekbewusmakingsprogramme in die bevordering van 'n woordeboekkultuur te beklemtoon.

Sleutelwoorde: WOORDEBOEKGEBRUIKSTRATEGIEË, WOORDEBOEKBEWUSMAKINGSPROGRAM, EKSPLISIETE EN GEÏNTEGREERDE STRATEGIE-ONDERRIG, WOORDEBOEKKULTUUR, CALLA, STRATEGIEGEBASEERDE ONDERRIG, NASLAANSTRATEGIEË, LEMMATISERINGSTRATEGIEË

1. Introduction

The dictionary is an important tool that may be used not only for looking up words, but also during writing. It is considered a valuable educational material that enhances literacy, the development of speech and language (Zarei and Gujjar 2012). Some scholars consider its use an effective learning strategy (Nation 1990, 2001; Gu and Johnson 1996; Scholfield 1997; Gu 2003), while others acknowledge its importance in vocabulary acquisition, reading or writing (Jackson 2002; Wingate 2002; Mohamad 2003; Fuertes-Olivera and Pérez Cabello de Alba 2012). The efficient use of a dictionary depends on the familiarity with dictionary using skills and knowledge of when using a specific dictionary or other tool, in other words dictionary culture (Gouws 2013), the reference skills (Hartmann and James 1998) and the dictionary use strategies (DUS) (Gavriilidou 2013), which refer to efficient dictionary users' decisions, behaviours and techniques regarding the internal processes they adopt, in order to perform successful dictionary searches. Previous research (Gavriilidou and Konstantinidou 2021) has shown that DUS are objective, observable, discoverable, amendable and teachable. Taking it as a given that DUS are teachable and building on previous research, the purpose of this study is to answer the question of which type of dictionary awareness program would be more beneficial for raising DUS of dictionary users. To do so, we investigated the effect of an explicit and integrated to language course dictionary familiarization program

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1882 (Article)

on primary school pupil's DUS. This article reports results of this investigation, starting with a literature review focusing on the construct of DUS and its inclusion in the recently reformed curricula for Greek Language Teaching in elementary and secondary schools in Greece (Magoula et al. 2022) and Cyprus (Mitsiaki 2020), followed by the research aims and hypotheses, methodology, results, discussion and conclusion of the study.

2. Literature review

2.1 Dictionary use strategies

Gavriilidou (2013) and Gavriilidou and Konstantinidou (2021) provide a comprehensive framework of DUS, that outlines the techniques and behaviors employed during dictionary look-ups. Gavriilidou (ibid.) explores the relationships between various variables, such as the task at hand, the type of dictionary used, and personal characteristics of dictionary users. The objective is twofold: to explain the complexity of dictionary use strategies and to offer a practical understanding of how these strategies contribute to the success or failure of dictionary searches.

Gavriilidou (2013) also aims to elucidate the actions taken by individual dictionary users to effectively complete their look-ups and to predict the role of DUS in improving look-up outcomes. For paper dictionaries, she (ibid.) classifies DUS into four categories: (1) Dictionary awareness strategies: These involve a critical awareness of the value and limitations of the dictionary, as well as an understanding of when and why to use a dictionary in specific circumstances; (2) Dictionary selection strategies: These enable users to choose an appropriate dictionary based on the problem they need to solve, ensuring familiarity with their chosen dictionary; (3) Lemmatization strategies: These assist dictionary users in finding the citation form of inflected words encountered in a text. Users rely on morphological indicators of the unknown word to make hypotheses about its look-up form. This category also encompasses skills in alphabetical sequencing; (4) Look-up strategies: These strategies facilitate the localization of the correct section of the entry where various meanings of a polysemous word form are included (for a detailed classification and definitions of DUS, see Gavriilidou and Konstantinidou 2021). This theory was extended in Mavrommatidou et al. (2019) and Gavriilidou et al. (2020) to cover digital dictionary use strategies.

DUS are problem-oriented as they are closely tied to specific learning tasks in language learning. They are action-based, requiring users to undertake specific actions to ensure successful word look-ups. Moreover, DUS are teachable and their selection is influenced by variables such as gender, motivation, learning style, educational and proficiency level, school type, task purpose, career orientation, and general reference skills (Gavriiidou et al. 2020; Gavriilidou and Konstantinidou 2021).

A growing body of research has emphasized the close relationship between

DUS and effective dictionary use (Chadjipapa et al. 2020; Gavriilidou et al. 2020). Effective dictionary use has, in turn, been found to correlate with successful performance in reading, writing, and vocabulary acquisition. Students who employ dictionary use strategies achieve more successful look-ups compared to those who do not strategize. Previous research has shown that effective dictionary users demonstrate better performance in reading comprehension (Knight 1994; Tono 1992; McCreary and Amacker 2006; Ma and Cheon 2018) and vocabulary acquisition (Hulstijn et al. 1996; Fraser 1999; Laufer 2000; Prichard 2008; Pousi 2010; Welker 2010; Hamilton 2012). A smaller number of studies have investigated the errors in dictionary use made by students during writing (Nesi and Meara 1994; Christianson 1997; Harvey and Yuill 1997; Hulstijn and Atkins 1998; Santos 2006; Elola, Rodríguez-García and Winfrey 2008). Additionally, some researchers (Harvey and Yuill 1997; Chun 2004) have compared the use of monolingual and bilingual dictionaries during writing and found a significant impact of dictionary use on the quality of the produced text. Overall, these studies highlight the importance of effective dictionary use and its impact on various language skills, including reading comprehension, vocabulary acquisition, and writing quality.

Unfortunately, many dictionary users are unaware of the complexity of their DUS. Alternatively, they may not develop sufficient mastery of the strategy repertoire independently, hindering their ability to conduct successful searches. Therefore, systematic training is necessary to enhance users' awareness and proficiency in employing a broad range of DUS for any task that requires the use of a dictionary. The question at hand is determining the most effective type of instruction for increasing students' awareness of DUS.

2.2 How to teach DUS?

A recent reform in the education system in Greece and Cyprus led national policy makers to acknowledge the value of cultivating dictionary culture and the importance of including dictionary training in classroom. This policy is mirrored in two national curricula recently compiled for teaching Greek as a second language in Cyprus (Mitsiaki 2020) and Greek as first language in Greece (Magoula et al. 2022: 15) which highlight every dimension of vocabulary knowledge, by promoting the creative use of dictionaries, both print and digital, and providing, for this purpose, targeted learning activities. As can be seen, from the learning outcomes/can do statements cited below, in these curricula dictionary use is closely connected to vocabulary acquisition:

> *Upon successful completion students will be able:*
> - *To get in touch with the children's dictionary.*
> - *To identify words that acquire special meaning in the context or identify words that are important in the context of the specific vocabulary.*
> - *To compare the content of entries from different types of printed and print media electronic dictionaries.*

- *To identify the thematic vocabulary of the texts in the teaching unit.*
- *To know how to search for words in printed or electronic (school) dictionaries school dictionaries.*
- *To extend their vocabulary with less frequent words by following the vocabulary of the school.*
- *To use dictionaries in order to verify their assumptions, to improve their production of spoken and written language.*
- *To identify the relationships of words based on their meaning: synonym or pronoun.*
- *To link their lexical choices to different levels of style and varying communication contexts.*
- *To form compound words based on thematic vocabulary.*
- *To create word families based on similar subject matter and focusing on meaning.*
- *Distinguish etymologica*lly related words.

(Mitsiaki 2020; Magoula et al. 2022)

Another relevant initiative in the Greek setting is the community-based *Curriculum for teaching Greek as a heritage language: a framework for teachers* (Gavriilidou and Mitsiaki 2022), compiled to be used as a framework for systematizing Greek heritage language teaching and testing in the USA with the purpose to empower Greek heritage speakers from pre-school to high school, so that, as teenagers, they will have gained a good knowledge of the varieties of Greek, of basic academic skills, and familiarized themselves with the Greek culture. In the four syllabi of the curriculum, dictionary awareness activities lead students to compile their own personal dictionary with the words they don't know so that they self-regulate their vocabulary learning and also train them to use efficiently DUS such as inferencing, self-monitoring, self-evaluation during receptive or productive use.

All the efforts described above promote a strategy-based instruction model in training DUS. Strategy training is defined here as any pedagogical approach and set of activities which provide language teachers with what they need to support dictionary users in enhancing their DUS by focusing on readily operationalizable strategies to be adopted and used by them to develop their reference skills, to improve particular task performance, or both.

DUS instruction is held explicitly, integrated into the language course content. This theoretical choice was based on previous literature which found, on the one hand, that explicit instruction is more effective because it cultivates students' metacognition by helping them reflect on their own learning and thinking (Anderson 2002; Chamot 2005; Sarafianou and Gavriilidou 2015) and, on the other, that "explicit teaching of DUS results in appropriate knowledge and skill development to successfully use a dictionary, raises the independence and confidence of students as dictionary users, increases their motivation to use a dictionary, which may be negatively affected by unsuccessful look ups, and develops their awareness of the positive strategies to be adopted while navigating in dictionary entries." (Gavriilidou and Konstantinidou 2021: 6). Furthermore, the integration of DUS instruction into the language course con-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1882 (Article)

tent helps dictionary users realize the usefulness of DUS used in connection with specific activities (reading, writing, listening, etc.), which facilitates retention. Students experience the advantages of systematically applying DUS to perform successful dictionary look-ups while engaging in different tasks during language learning. In addition, they have opportunities to share their own preferred DUS with the other dictionary users in the class and to increase their strategy repertoires within the context of the typical language tasks they are asked to engage in.

3. Aim and research questions

Numerous researchers have emphasized the importance of teaching effective dictionary use and have suggested training as a means to enhance users' reference skills and DUS (cf. e.g. Herbst and Stein 1987; Gavriilidou 2017). Additionally, several studies have examined the impact of dictionary awareness programs on the development of reference skills and the overall improvement of dictionary use effectiveness (Głowacka 2001; Carduner 2003; Chi 2003; Gavriilidou and Sfyroera 2004; Gavriilidou 2017). Considering the need for more focused studies investigating teacher-led approaches with diverse language students in different learning contexts worldwide (Cohen and Macaro 2007; Plonsky 2011), our primary objective is to investigate whether an explicit and integrated dictionary awareness program can lead to changes in self-reported strategy use among primary school Greek students.

By conducting this study, we aim to contribute to existing research and address the gap in the literature regarding the impact of teacher-led dictionary awareness programs on dictionary use strategies.

The research questions guiding this study are:

1. What is the level of strategic dictionary use of the participants before the implementation of the program? Previous research (Bensoussan et al. 1984; Neubach and Cohen 1988; Beech 2004; Chadjipapa et al. 2020) suggests that the participants in the study are anticipated to exhibit moderate engagement with DUS.
2. Does a comprehensive strategy instruction program influence the self-reported dictionary strategy use of upper elementary Greek students? While a significant portion of existing literature on program implementations focuses on their impact on overall language proficiency or distinct linguistic skills (Sengupta 2000; Macaro 2001; Carrier 2003), several studies have explored the efficacy of these programs in enhancing reference skills, thereby improving dictionary use (Głowacka 2001; Carduner 2003; Chi 2003; Gavriilidou 2017).

Through these research objectives, this study seeks to deepen our understanding of the impact of an explicit and integrated strategy instruction program on

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1882 (Article)

the self-reported use of DUS among upper elementary Greek students.

4. Methods

4.1 Research design

The study employed a quasi-experimental design, specifically a "pre-test-post-test control-group design." This design, rooted in the quantitative paradigm, leverages the sample survey technique, apt for gathering data via structured questionnaires to discern opinions, perceptions, attitudes, and beliefs. This methodological choice is congruent with the study's overarching aim, specific objectives, and research inquiries.

4.2 Participants

The study involved 150 students, roughly balanced between males (49.3%) and females (50.7%). These participants hailed from two distinct school types (mainstream and intercultural) in Komotini and Ierapetra of Crete (Greece). Convenience sampling was the method of choice for participant selection. The students were from the 5th and 6th grades of elementary school. The control group comprised two 6th-grade classes and one 5th-grade class, each with 25 students (75 in total). Conversely, the experimental group had two 5th-grade classes and one 6th-grade class, each with 25 students (75 in total). Both groups underwent diagnostic (pre-test) and evaluative (post-test) assessments simultaneously. However, only the experimental group experienced the teaching implementation. Gender distribution was nearly equal across both groups. Specifically, the experimental group comprised 38 males (25.3%) and 37 females (24.7%), whereas the control group consisted of 36 males (24.0%) and 39 females (26.0%).

4.3 Procedure

The research unfolded in three phases. Initially, all the participants detailed their typical strategies during dictionary look-ups using the Strategy Inventory for Dictionary Use (SIDU) (Gavriilidou 2013). Subsequently, the experimental group, which consisted of three sections, underwent an experimental dictionary awareness program (detailed in section 4.5), spanning four weeks (2 hours daily). The control group did not undergo any specific training. During the concluding phase, both groups revisited the SIDU immediately following the program's completion. The retention measure, originally planned for three months after the intervention, was expedited due to the lockdown enforced amid the COVID-19 pandemic. The pre-test responses were compared with the post-test ones to identify potential differences in DUS between the two measurements.

Before the students completed the questionnaire, they were informed about its purpose and content. This information was provided either by the researcher or by the responsible teacher of the experimental group. In the two sections where the researcher was not present, the teachers received clear instructions on how to present the research's purpose and how to administer the questionnaire. The study was approved by the Ethics Committee of the Department of Greek of Democritus University of Thrace. Written consent was obtained from the legal guardians of minors.

Throughout the process of completing the questionnaires, either the researcher or the assigned teacher was present to offer clarifications whenever required by the students. The allocated time for each section of the experimental group to complete the questionnaire was one school hour, equivalent to 45 minutes, which proved sufficient for smooth completion. None of the sections exceeded half an hour to complete the questionnaire, and no delays occurred due to time-consuming explanations.

4.4 Instrumentation

The SIDU (Gavriilidou 2013) was the primary tool to gauge dictionary strategy usage before and after the implementation of the program. This 36-item self-report questionnaire delves into dictionary strategy utilization across four strategy categories: (a) Awareness strategies (Questions 1–14), e.g., "6. I use a dictionary to find the origin of a word." (b) Selection Strategies (Questions 15–21), e.g., "18. I know what an etymological dictionary is and what it is used for." (c) Lemmatization Strategies (Questions 22–29), e.g., "25. When I can't locate a proverb or a set phrase in the entry where I thought I would find it, I begin a new search." (d) Look-up Strategies (Questions 30–36), e.g., "35. When I find the word that I was searching for, I return to the text to confirm that the word matches the context" (Gavriilidou 2013).

Developed meticulously, its validity has been previously established (Gavriilidou 2013). Every item of the instrument was reviewed by multiple experts for clarity and content validity. During pilot measures, SIDU was rigorously assessed for social desirability response bias, revealing a range of DUS among respondents. Importantly, these strategies were not concentrated at either the extremely high or median levels, providing evidence that SIDU did not produce socially desirable results.

The SIDU employs a five-point Likert scale to measure the frequency of strategy use (e.g. "never or almost never" was coded 1, "usually never" was coded 2, "sometimes" was coded 3, "usually" was coded 4, and "always or almost always" was coded 5). The English version of the instrument can be found in **Appendix A**.

Compared to other data collection protocols utilized in the study of reference skills or dictionary use, self-report composite rating scales like SIDU offer notable advantages. They are simple and swift to administer, providing a

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1882 (Article)

broad assessment of each student's typical self-reported dictionary use strategies. Additionally, they facilitate the collection of data from large samples in a cost-effective and time-efficient manner. Moreover, self-report instruments are commonly employed in the study of Language Learning Strategies.

While alternative methods of investigating dictionary use also yield valuable insights, they possess inherent limitations. For instance, observation during dictionary look-up is straightforward but fails to capture data on unobservable dictionary use strategies. Interviews offer personalized information but are labor-intensive. Think-aloud protocols provide detailed insights, contingent upon users' willingness and ability to articulate their internal behaviors. Eye-tracking methods offer meticulous information but are typically limited to one-to-one settings, time-consuming, unsuitable for large-scale studies, such as the one presented here, and lack summative capabilities across students for group analysis (for a detailed report on data collection methods for the investigation of users' dictionary consultations see Tono 2001 and Lew et al. 2013).

Consequently, SIDU was chosen as the primary instrument for reliably gathering large-scale self-reported data, which can be complemented and triangulated with data obtained through other data collection methods.

4.5 The dictionary awareness program

The program, executed with the experimental group, adheres to Cognitive Academic Language Learning Approach (CALLA) principles and unfolds through five phases: (a) **Preparation:** Here, students discuss their dictionary usage habits and cultivate metacognitive understanding of the connection between DUS and effective look-ups. (b) **Presentation:** In this phase, teachers demonstrate DUS, elucidating their features and applications. (c) **Practice/ Scaffolding:** Students engage in exercises involving the DUS discussed, within genuine learning contexts. (d) **Self-evaluation:** Students reflect on the application of their DUS and the associated metacognitive insights. (e) **Expansion:** Students transpose their chosen DUS across varied contexts (Gavriilidou and Konstantinidou 2021). The program is both explicit and integrated. Throughout its course, students are instructed on the optimal contexts and reasons for employing specific dictionary strategies. This approach ensures that learners can autonomously rectify their errors during their educational journey (Larsen-Freeman 2000; Richards and Rodgers 2007). Additionally, the program embraces differentiated instruction and integrates intercultural learning dimensions. It offers tailored activities catering to users from diverse linguistic or cultural backgrounds. Spanning 12 units, the program offers focused instruction on DUS in a printed format, tailored for 5th and 6th-grade students. It incorporates exercises that enhance vocabulary strategies, drawing from relevant sections of the prescribed textbook. The program aligns with the curriculum set by the Ministry of Education, themes in the 5th and 6th-grade Greek textbooks, and accompanying teacher resources. While the implementation of the program spanned four weeks (equating to 40 school hours), its duration can be adapted to suit the unique requirements, levels, and interests of individual classes.

4.6 Data analysis

The internal consistency reliability of the four types of DUS was assessed using Cronbach's alpha. To validate the factor structure of the SIDU questionnaire, a Confirmatory Factor Analysis (CFA) was conducted on the proposed four-strategy model. The model's goodness-of-fit was assessed using several fit indices: Comparative FitIndex (CFI), Tucker-Lewis Index (TLI), Standardized Root Mean Square Residual (SRMR), Root Mean Square Error of Approximation (RMSEA). The CFI and TLI values, both exceeding 0.90, along with the SRMR value below 0.08 and the RMSEA value below 0.05, collectively indicate a satisfactory fit of the model to the data. Subsequent analyses involved computing the average scores for each strategy type based on individual items, to evaluate the level of strategic dictionary use before and after program implementation. To investigate the effects of the program on the frequency of strategy use, a two-way repeated measures Analysis of Variance (ANOVA) was conducted. This analysis was chosen to account for the two independent variables: group type (Experimental vs. Control) and time (Pre-test vs. Post-test). The dependent variable was the frequency of strategy use. By employing this statistical approach, we aimed to discern any significant interactions between the group type and time, which would indicate the program's differential impact on strategy usage frequency across the two groups. Eta squared (η^2) was used as a measure of effect size to quantify the magnitude of the observed effects. Conventionally, values of 0.01, 0.06, and 0.14 are considered to represent small, medium, and large effect sizes, respectively (Cohen 1988). All statistical analyses were executed in R, utilizing the aov function from the base package and the cfa function from the lavaan package (Rosseel 2012).

5. Results

The four-strategy CFA model demonstrated a satisfactory fit. Fit indices included CFI (0.921), TLI (0.918), SRMR (0.0555), and RMSEA (0.0421; 90% CI: 0.0349–0.0491). These values indicate a strong and close fit of the model to the observed data. Factor loadings for all categories were significant at $p < 0.001$: awareness (0.43–0.69), selection (0.45–0.73), lemmatization (0.51–0.74), and look-up (0.57–0.74). For internal consistency, the Cronbach's alpha values were 0.87 for awareness, 0.77 for selection, 0.82 for lemmatization' and 0.84 for look-up strategies, indicating the SIDU questionnaire's reliable measurement of each strategy type.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1882 (Article)

Table 1: Comparison of strategy use in experimental and control groups before and after the program's implementation

		Pre-test		Post-test	
Strategy type	**Group**	**Mean**	**SD**	**Mean**	**SD**
Awareness	Experimental	2.83	0.79	3.60	0.64
	Control	2.33	0.65	2.48	0.68
Selection	Experimental	2.97	0.90	3.66	0.80
	Control	2.56	0.86	2.63	0.76
Lemmatization	Experimental	3.22	0.98	3.77	0.80
	Control	2.80	0.94	2.77	0.84
Look-up	Experimental	3.44	0.93	4.05	0.62
	Control	3.12	0.85	3.17	0.79
Overall strategy use	Experimental	3.07	0.71	3.75	0.59
	Control	2.65	0.66	2.72	0.62

As illustrated in Table 1, the mean scores and standard deviations for each strategy type — awareness, selection, lemmatization, and look-up — were compared between the experimental and control groups both before and after the program's implementation.

Overall strategy use: Table 1 and Figure 1 present the overall dictionary use strategies employed by students in each group, both before and after the program's implementation. Prior to the implementation, an ANOVA analysis found no significant difference in strategy frequency between the experimental and control groups (Mean Difference = 0.426, $p = 0.07$), suggesting that both groups were on a similar footing before the program's onset. However, a significant interaction emerged between the group and the time of measurement ($F(1, 148) = 35.997$, $p < 0.001$, $\eta^2 = 0.196$). This effect size, as denoted by the η^2 value, implies that nearly 19.6% of the variance in overall strategy use can be attributed to the combined influence of the group and the time of measurement, which is indicative of a large effect size. Before the program's implementation, students in the control group reported employing the strategies with a low to moderate frequency — a pattern that remained relatively unchanged after program's implementation (Mean Difference = 0.074, $p = 0.298$). Conversely, the experimental group students exhibited a marked increase in the frequency of overall strategy use after the implementation of the program (Mean Difference = 0.676, $p < 0.001$). This is also illustrated in Figure 1. Such findings emphasize the program's potential effectiveness in bolstering strategic dictionary use, especially within the experimental group.

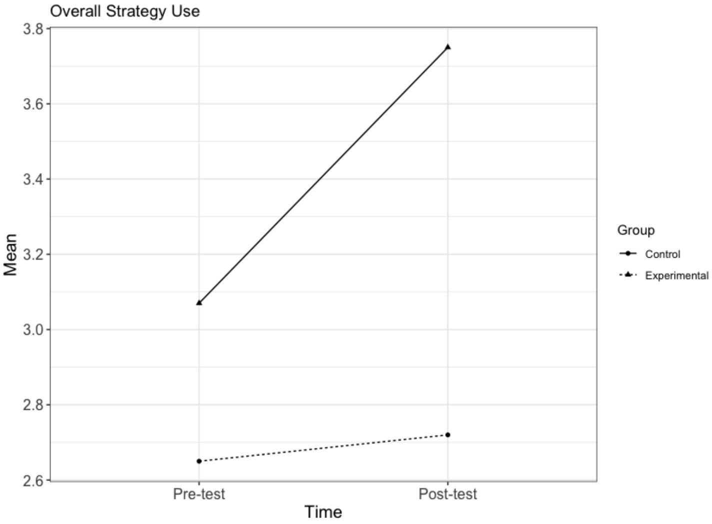


Figure 1: Overall strategy use before and after the implementation of the program in experimental and control groups

Awareness strategy use: As depicted in Table 1 and Figure 2, students' familiarity with dictionary use contexts was assessed before and after the implementation of the program. Initial comparisons revealed no significant difference in strategy frequency between the experimental and control groups (Mean Difference = 0.498, $p = 0.06$). However, a significant interaction between group and time of measurement was observed ($F(1, 148) = 30.506$, $p < 0.001$, $\eta^2 = 0.171$). The control group's use of awareness strategies remained consistent after the implementation of the program (Mean Difference = 0.146, p = 0.071), while the experimental group showed a notable increase (Mean Difference = 0.771, $p < 0.001$).

Selection strategy use: Both groups were initially comparable in selection strategy use with a Mean Difference of 0.410, $p = 0.07$ (see Table 1 and Figure 3). A significant interaction between group and time was noted ($F(1, 148) = 21.101$, $p < 0.001$, $\eta^2 = 0.125$). The control group maintained their strategy use after the implementation of the program (Mean Difference = 0.065, $p = 0.500$), whereas the experimental group reported increased use (Mean Difference = 0.688, $p < 0.001$).

Lemmatization strategy use: As presented in Table 1 and Figure 4, lemmatization strategies were consistent across groups before the implementation of the program (Mean Difference = 0.417, $p = 0.07$). A significant interaction was detected after the implementation of the program ($F(1, 148) = 19.744$, $p < 0.001$, $\eta^2 = 0.117$). The control group's strategy use remained stable (Mean Difference = 0.029, $p =$

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1882 (Article)

0.756), while the experimental group showed a marked increase (Mean Difference = 0.55, *p*< 0.001).

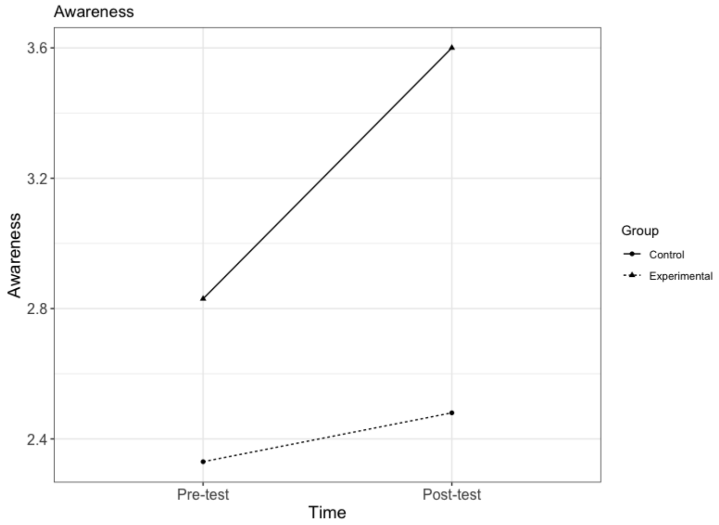


Figure 2: Awareness strategy use before and after the implementation of the program in experimental and control group

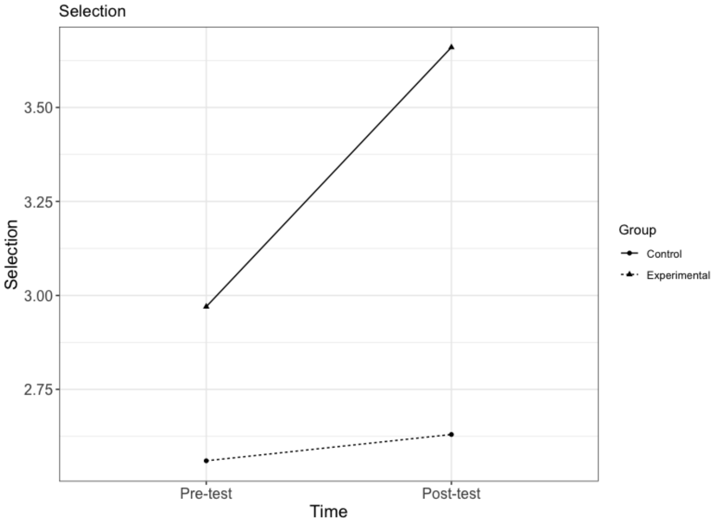


Figure 3: Selection strategy use before and after the implementation of the program in experimental and control groups

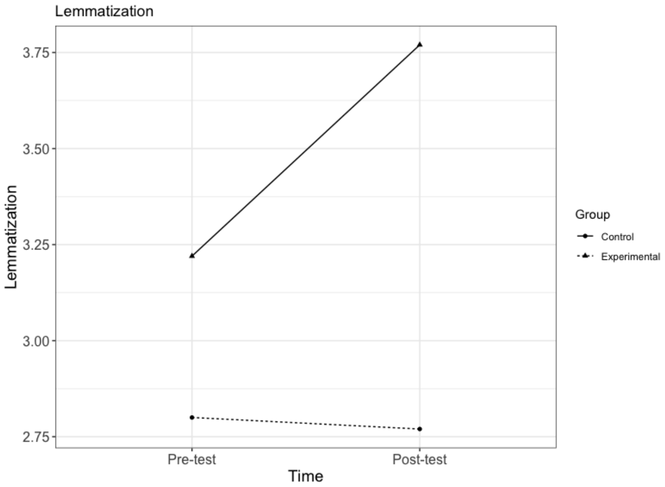


Figure 4: Lemmatization strategy use before and after the implementation of the program in experimental and control groups

Look-up strategy use: Table 1 and Figure 5 detail students' look-up strategies. Both groups were comparable in strategy frequency before the implementation of the program (Mean Difference = 0.320, p = 0.12). A significant interaction was observed after the implementation of the program ($F(1, 148) = 19.612$, $p < 0.001$, $\eta^2 = 0.117$). The control group's strategy use remained consistent (Mean Difference = 0.047, p = 0.605), while the experimental group reported a significant uptick (Mean Difference = 0.61, $p < 0.001$).

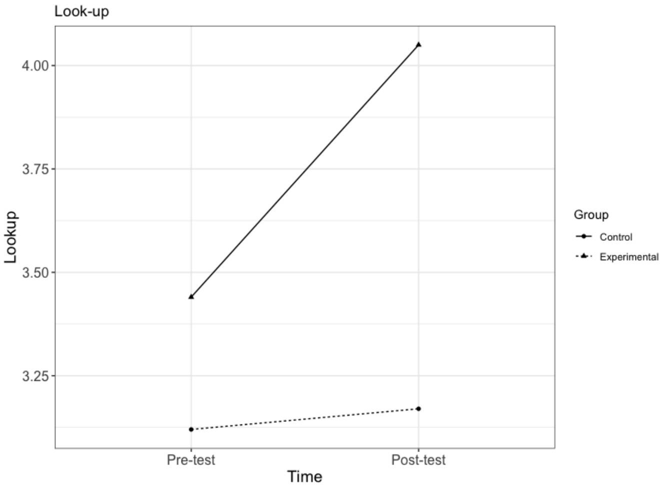


Figure 5: Look-up strategy use before and after the implementation of the program in experimental and control groups

6. Discussion

The first research question in this survey aimed to assess the level of strategic dictionary use among the participants in the study based on categories before the implementation of the program. Considering previous research (Bensoussan et al. 1984; Neubach and Cohen 1988; Beech 2004; Chadjipapa et al. 2020), it was hypothesized that the participants would exhibit a moderate level of dictionary use strategies overall and within each category.

The statistical analyses conducted confirmed the initial hypotheses, as learners reported utilizing dictionary use strategies overall and within each category to a moderate degree. Therefore, it can be concluded that the users in the sample cannot be classified as "strategic" dictionary users, as they demonstrate only a moderate extent of strategy usage. This moderate level of use may be attributed to the non-conscious use of strategies by the survey participants or the lack of systematic and organized dictionary use in Greek general and intercultural schools, which lack targeted activities and appropriate motivation from teachers.

These findings underscore the need for increased awareness among teachers regarding the importance of supporting learners in becoming more proficient users of dictionaries. Enhancing this could be accomplished by engaging teachers in tailored professional development sessions focused on a strategic dictionary use program designed to cater to their unique requirements and challenges. These findings also suggest that further efforts are needed to enhance the moderate use of all types of DUS, foster a dictionary culture among elementary and secondary pupils, and increase awareness of the benefits of dictionary use and its potential to improve students' lexical knowledge. Consequently, continuous in-service training is necessary for teachers to develop expertise and effectively incorporate DUS into the Greek educational setting.

The second research question examined the effect of the dictionary awareness program on dictionary use among the students. The results indicate a significant effect of the program, with all students in the experimental group demonstrating an increase in the use of dictionary use strategies overall and within individual categories.

While before the implementation of the program, the students in the sample exhibited a moderate level of dictionary use strategies, after the implementation of the program, students in the experimental group reported a significantly higher use of strategies overall compared to the control group. The frequency of dictionary use within the control group exhibited no significant alterations following the program's implementation, maintaining a consistently low to moderate level comparable to pre-program levels.

Specifically, in the awareness strategies, selection strategies, lemmatization strategies, and look-up strategies, the two groups (experimental and control), which were considered equivalent before the implementation of the program, showed significant differences after its implementation. Before the implementation of the program, the control group reported low to moderate usage of the strategies, and this frequency did not change significantly after the implementation of the program. In contrast, the experimental group reported a significantly higher use of all categories of dictionary use strategies after the implementation of the program, with the most notable increase observed in the awareness strategies and selection strategies, which had the lowest rates of use before the implementation of the program.

The increase in the frequency of strategy use after the implementation of the program is a positive indication of the impact of the teaching approach in the context of Greek language teaching in primary schools. However, it is important to note that the success of this approach relies on redefining the role of the teacher and implementing student-centered methods that promote autonomy in dictionary use.

Overall, these findings provide additional support for the "teachability" of dictionary use strategies and skills, suggesting that well-designed dictionary awareness programs can heighten awareness and cultivate a dictionary culture. It further supports the claim that explicit strategy instruction can lead to increased dictionary use. Additionally, the effective implementation of the program indicates that a direct and clear presentation of DUS is more likely to yield success than an implicit approach and contribute to the development of autonomous learners in vocabulary acquisition.

7. Conclusions, limitations and further research

This study investigated the effectiveness of a dictionary awareness program that focused on explicit and integrated strategy training for primary school students attending mainstream and intercultural schools in Greece. The findings revealed a moderate degree of dictionary usage, as reported by the students in the sample, indicating an incomplete integration of dictionaries as reference tools in the educational process. However, following the program's implementation, a notable surge in the overall adoption of dictionary use strategies was evident. This positive outcome provides encouraging evidence of the specific teaching approach's impact, aligning with the long-term goal of cultivating strategically autonomous learners proficient in using dictionaries.

This study does, however, have a few limitations. First of all, the current study was based on a quantitative research design involving a questionnaire survey. The combination of quantitative and qualitative methods could have reinforced the internal validity of the study and could have provided further insights regarding the learners' ability in choosing appropriate strategies. Furthermore, the assignment of subjects to the experimental and control groups relied on the researcher's convenient accessibility rather than random sampling, thereby compromising the external validity of the research and constraining the generalizability of the findings. Third, in light of the lockdown enforced amid the COVID-19 pandemic, the retention measure initially scheduled for three months post-inter-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1882 (Article)

vention was expedited. Finally, although the indications of the results after the implementation of the program are very encouraging, the short duration (4 weeks) limits the possibility of a more complete assessment of the program.

A recommended approach for future research could involve replicating the survey while considering various age demographics and incorporating a retention measure to assess the sustainability of the observed effects over time. It would also be of particular interest to study the effect of further parameters on the strategic use of the dictionary, such as the socio-economic context, the performance of the student's performance, motivation and learning trajectory, since according to the literature, these variables influence the use of learning strategies (Lan and Oxford 2003; Chamot and Keatley 2004; Oxford et al. 2004; Gavriilidou and Petrogiannis 2016; Gavriilidou et al. 2017). A defining suggestion for future research could involve expanding the methodology to include qualitative techniques, such as observational studies, which could unveil nuanced insights into the efficacy and utilization of strategies. This approach would offer a deeper understanding of the topic by capturing real-time behaviors and contextual factors that quantitative measures may overlook. Lastly, an additional research recommendation would be to investigate the correlation between the dictionary strategies employed by teachers and those utilized by students. This comparative analysis could offer valuable insights into the dynamics of strategy transmission and adoption within educational settings, shedding light on potential influences and reciprocal effects between educators and learners.

References

Anderson, N.J. 2002. *The Role of Metacognition in Second Language Teaching and Learning.* Eric Digest. Washington, DC: ERIC Clearinghouse on Languages and Linguistics.

Beech, J.R. 2004. Using a Dictionary: Its Influence on Children's Reading, Spelling, and Phonology. *Reading Psychology* 25(1): 19-36.

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. *Readings in a Foreign Language* 2(2): 262-276.

Carduner, J. 2003. Productive Dictionary Skills Training: What Do Language Learners Find Useful? *Language Learning Journal* 28: 70-76.

Carrier, K.A. 2003. Improving High School English Language Learners' Second Language Listening through Strategy Instruction. *Bilingual Research Journal* 27(3): 383-408.

Chadjipapa, E., Z. Gavriilidou, A. Markos and A. Mylonopoulos. 2020. The Effect of Gender and Educational Level on Dictionary Use Strategies Adopted by Upper-elementary and Lower-secondary Students Attending Greek Schools. *International Journal of Lexicography* 33(4): 443-462.

Chamot, A.U. 2005. Language Learning Strategy Instruction: Current Issues and Research. *Annual Review of Applied Linguistics* 25: 112-130.

Chamot, A.U. and C.W. Keatley. 2004. Learning Strategies of Students of Less Commonly Taught Languages. Paper presented at the 2004 Annual Meeting of the American Educational Research Association, San Diego, CA, 12–16 April 2004.

Chi, A.M.L. 2003. *An Empirical Study of the Efficacy of Integrating the Teaching of Dictionary Use into a Tertiary English Curriculum in Hong Kong.* Hong Kong: Language Centre, Hong Kong University of Science and Technology.

Christianson, K. 1997. Dictionary Use by EFL Writers: What Really Happens? *Journal of Second Language Writing* 6(1): 23-43.

Chun, Y.V. 2004. EFL Learners' Use of Print and Online Dictionaries in L1 and L2 Writing Processes. *Multimedia-Assisted Language Learning* 7(1): 9-35.

Cohen, A.D. and E. Macaro. 2007. *Language Learner Strategies.* Oxford: Oxford University Press.

Cohen, J. 1988. *Statistical Power Analysis for the Behavioral Sciences.* Second edition. Hillsdale, New Jersey: Lawrence Erlbaum.

Elola, I., V. Rodríguez-García and K. Winfrey. 2008. Dictionary Use and Vocabulary Choices in L2 Writing. *Estudios de Lingüística Inglesa Aplicada* 8: 63-89. Retrieved from: http://institucional.us.es/revistas/elia/8/6.%20elola%20def.pdf

Fraser, C.A. 1999. The Role of Consulting a Dictionary in Reading and Vocabulary Learning. *Canadian Journal of Applied Linguistics* 2(1–2): 73-89.

Fuertes-Olivera, P.A. and M.B. Pérez Cabello de Alba. 2012. Online Dictionaries and the Teaching/Learning of English in the Expanding Circle. *International Journal of English Studies* (IJES) 12(1): 147-166.

Gavriilidou, Z. 2013. Development and Validation of the *Strategy Inventory for Dictionary Use* (S.I.D.U.). *International Journal of Lexicography* 26(2): 135-153.

Gavriilidou, Z. 2014. Translation, Cultural Adaptation and Preliminary Psychometric Evaluation of the English Version of "Strategy Inventory for Dictionary Use" (S.I.D.U). Abel, A., Ch. Vettori and N. Ralli (Eds.). 2014. *Proceedings of the XVI Euralex International Congress: The User in Focu*s, *EURALEX 2014, Bolzano/Bozen, Italy, July 15–19, 2014:* 225-235. Bolzano/Bozen: Institute for Specialised Communication and Multilingualism. http://www.euralex.org/elx_proceedings/Euralex2014/euralex_2014_015_p_225.pdf

Gavriilidou, Z. 2017. The Effect of an Intervention Programme on Improving Electronic Dictionary Reference Skills of Students Attending Secondary Schools in Greece. Oral Presentation at Elex 2017, Electronic Lexicography in the 21st Century: Lexicography from Scratch, Leiden, The Netherlands, 19–21 September 2017.

Gavriilidou, Z. and E. Konstantinidou. 2021. The Design of an Explicit and Integrated Intervention Program for Pupils Aged 10–12 with the Aim to Promote Dictionary Culture and Strategies. Gavriilidou, Z., L. Mitits and A. Kiosses, S. (Eds.). 2021. *Proceedings of the XIX Euralex International Congress Alexandroupolis, Greece, 7–11 September 2021.* Volume 2: 735-745. Alexandroupolis: Democritus University of Thrace.

Gavriilidou, Z. and M. Mitsiaki. 2022. *Curriculum for Teaching Greek as a Heritage Language: A Framework for Teachers*. Komotini: 2KProject. Available at: https://synmorphose.gr/index.php/el/publications-gr/syllabi-menu-gr

Gavriilidou, Z. and K. Petrogiannis. 2016. Language Learning Strategy Use of English FL Learners in Greek Schools: The Role of School Type and Educational Level. *International Journal of Research Studies in Language Learning* 5(4): 67-81

Gavriilidou, Z. and M. Sfyroera. 2004. Elaboration et usage d'un dictionnaire par des enfants d'Age pré-scolaire. Braasch, A. and C. Povlsen (Eds.). 2002. *Proceedings of the Tenth EURALEX International Congress, EURALEX 2002, Copenhagen, Denmark, August 13–17, 2002. Volume 1:* Copenhagen: Center for Sprogteknologi, Københavns Universitet.

Gavriilidou, Z., S. Mavrommatidou and A. Markos. 2020. The Effect of Gender, Age and Career Orientation on Digital Dictionary Use Strategies. *International Journal of Research Studies in Education* 9(6): 63-76.

Gavriilidou, Z., K. Petrogiannis, M. Platsidou and A. Psaltou-Joycey. 2017. *Language Learning Strategies: Theoretical Issues and Applied Perspectives.* Kavala: Saita Publications.

Głowacka, W. 2001. *Difficulties with Understanding Dictionary Labels Experienced by Polish Learners of English Using Bilingual Dictionaries.* M.A. Dissertation, Adam Mickiewicz University, Poznań, Poland.

Gouws, R.H. 2013. Establishing and Developing a Dictionary Culture for Specialised Lexicography. Jesensěk, V. (Ed.). 2013. *Specialised Lexicography. Print and Digital, Specialised Dictionaries, Databases:* 51-62. Berlin/Boston: De Gruyter.

Gu, P.Y. 2003. Vocabulary Learning in a Second Language: Person, Task, Context and Strategies. *Tesl-Ej 7*(2): 1-25.

Gu, Y. and R.K. Johnson. 1996. Vocabulary Learning Strategies and Language Learning Outcomes. *Language Learning* 46(4): 643-679.

Hamilton, H. 2012. The Efficacy of Dictionary Use while Reading for Learning New Words. *American Annals of the Deaf* 157(4): 358-372.

Harvey, K. and D. Yuill. 1997. A Study of the Use of a Monolingual Pedagogical Dictionary by Learners of English Engaged in Writing. *Applied Linguistics* 18(3): 253-278.

Herbst, T. and G. Stein. 1987. Dictionary-using Skills: A Plea for a New Orientation in Language Teaching. Cowie, A.P. (Ed.). 1987. *The Dictionary and the Language Learner. Papers from the EURALEX Seminar at the University of Leeds, 1–3 April 1985:* 115-127. Tübingen: Max Niemeyer.

Hulstijn, J.H. and B.T.S. Atkins. 1998. Empirical Research on Dictionary Use in Foreign Language Learning: Survey and Discussion. Atkins, B.T.S. (Ed.). 1998. *Using L2 Dictionaries: Studies of Dictionary Use by Language Learners and Translators:* 7-19. Tübingen: Max Niemeyer.

Hulstijn, J.H., M. Hollander and T. Greidanus. 1996. Incidental Vocabulary Learning by Advanced Foreign Language Students: The Influence of Marginal Glosses, Dictionary Use, and Reoccurrence of Unknown Words. *The Modern Language Journal* 80(3): 327-339.

Jackson, H. 2002. *Lexicography: An Introduction.* London/New York: Routledge.

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://www.jstor.org/stable/330108

Lan, R. and R. Oxford. 2003. Language Learning Strategy Profile of Elementary School Students in Taiwan. *IRAL* 41(4): 339-379.

Larsen-Freeman, D. 2000. *Techniques and Principles in Language Teaching.* Second Edition. Oxford: Oxford University Press.

Laufer, B. 2000. Electronic Dictionaries and Incidental Vocabulary Acquisition: Does Technology Make a Difference? 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:* 849-854. Stuttgart: Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

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.

Ma, J.H. and H.J. Cheon. 2018. An Experimental Study of Dictionary Use on Vocabulary Learning and Reading Comprehension in Different Task Conditions. *International Journal of Lexicography* 31(1): 29-52
https://doi.org/10.1093/ijl/ecw037

Macaro, E. 2001. Analysing Student Teachers' Codeswitching in Foreign Language Classrooms: Theories and Decision Making. *The Modern Language Journal* 85(4): 531-548.

Magoula, E., N. Mitsis, D. Kanellopoulos, A. Mitsis, S. Samara, M. Oikonomaku, A. Tzanaki and I. Tryfiatis. 2022. *Curriculum for the Modern Greek Language Course* Institute of Educational Policy [In Greek].

Mavrommatidou, S., Z. Gavriilidou and A. Markos. 2019. Development and Validation of the Strategy Inventory for Electronic Dictionary Use (S.I.E.D.U.). *International Journal of Lexicography* 32(4): 393-410.
https://doi.org/10.1093/ijl/ecz015

McCreary, D.R. and E. Amacker. 2006. Experimental Research on College Students' Usage of Two Dictionaries: A Comparison of the *Merriam-Webster Collegiate Dictionary* and the *Macmillan English Dictionary for Advanced Learners*. Corino, E., C. Marello and C. Onesti (Eds.). 2006. *Proceedings XII EURALEX International Congress, Turin, Italy, September 6th–9th, 2006*: 871-885. Alessandria: Edizioni dell'Orso.

Mitsiaki, M. 2020. Detailed Curriculum for Greek as a Second Language. (Pre-primary, Primary, Secondary General, Secondary Technical and Vocational Education and Training of Cyprus). Ministry of Education, Nicosia, Cyprus. Electronic version [In Greek].
http://www.moec.gov.cy/analytika_programmata/analytika_programmata/ellinika_defteri_glossa/ap_analytiko_programma.pdf

Mohamad, N. 2003. *Penggunaan kamus Arab–Melayu di kalangan pelajar sekolah menengah* [The Use of Arabic–Malay Dictionary among Secondary School Students]. Unpublished M.A. thesis, Faculty of Language and Linguistic, University of Malaya.

Nation, I.S.P. 1990. *Teaching and Learning Vocabulary.* New York: Newbury House.

Nation, I.S.P. 2001. *Learning Vocabulary in Another Language.* Cambridge: Cambridge University Press.

Nesi, H. and P. Meara. 1994. Patterns of Misinterpretation in the Productive Use of EFL Dictionary Definitions. *System* 22(1): 1-15.

Neubach, A. and A.D. Cohen. 1988. Processing Strategies and Problems Encountered in the Use of Dictionaries. *Dictionaries. Journal of the Dictionary Society of North America* 10: 1-19. Retrieved from: http://dx.doi.org/10.1353/dic.1988.0018

Oxford, R., Y. Cho, S. Leung and H.J. Kim. 2004. Effect of the Presence and Difficulty of Task on Strategy Use: An Exploratory Study. *IRAL* 42: 1-47.

Plonsky, L. 2011. The Effectiveness of Second Language Strategy Instruction: A Meta-analysis. *Language Learning* 61(4): 993-1038.

Pousi, A. 2010. *Training in Dictionary Use: A Teaching Intervention in a 9th Grade EFL Classroom in Finland.* Bachelor's thesis, University of Jyväskylä, Finland.

Prichard, C. 2008. Evaluating L2 Readers' Vocabulary Strategies and Dictionary Use. *Reading in a Foreign Language* 20(2): 216-231.

Richards, J.C. and T.S. Rodgers. 2007. *Approaches and Methods in Language Teaching.* Cambridge: Cambridge University Press.

Rosseel, Y. 2012. lavaan: An R package for Structural Equation Modeling. *Journal of Statistical Software* 48(2): 1-36.

Santos, S. 2006. Dictionary Use in L2 Writing. *Memorias del II Foro Nacional de Estudios en Lenguas* (FONAEL 2006): 1-12.

Sarafianou, A. and Z. Gavriilidou. 2015. The Effect of Strategy-Based Instruction on Strategy Use by Upper-Secondary Greek Students of EFL. *Electronic Journal of Foreign Language Teaching* 12(1): 21-34. Available online at:
http://e-flt.nus.edu.sg/v12n12015/sarafianou.pdf

Scholfield, P. 1997. Vocabulary Reference Works in Foreign Language Learning. Schmitt, N. and M. McCarthy (Eds.). 1997. *Vocabulary: Description, Acquisition, and Pedagogy:* 279-302. Cambridge: Cambridge University Press.

Sengupta, S. 2000. An Investigation into the Effects of Revision Strategy Instruction on L2 Secondary School Learners. *System* 28(1): 97-113.

Tono, Y. 1992. The Effect of Menus on EFL Learners' Look-up Processes. *Lexikos* 2: 230-253.

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.

Welker, H.A. 2010. *Dictionary Use: A General Survey of Empirical Studies.* Brasilia: Author's Edition.

Wingate, U. 2002. *The Effectiveness of Different Learner Dictionaries. An Investigation into the Use of Dictionaries for Reading Comprehension by Intermediate Learners of German.* Lexicographica Series Maior 112. Tübingen: Niemeyer

Zarei, A.A. and A.A. Gujjar. 2012. The Contribution of Electronic and Paper Dictionaries to Iranian EFL Learner's Vocabulary Learning. *International Journal of Social Sciences & Education* 2(4): 628-634.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1882 (Article)

Appendix A: English version of S.I.D.U (Gavriilidou 2014)

Name (not surname):
Gender:
Date of birth:
Mother Tongue:
Career orientation:

This questionnaire will be used for research purposes and your contribution is very significant. Thank you for your help. Please read the following statements carefully and circle 1, 2, 3, 4 or 5 according to what is most true for you.

(1) Never or almost never true of me.
(2) Generally not true of me.
(3) Somewhat true of me.
(4) Generally true of me.
(5) Always true of me.

I use a dictionary to find the meaning of a word	1	2	3	4	5
I use a dictionary to find the spelling of a word	1	2	3	4	5
I use a dictionary to find synonyms	1	2	3	4	5
I use a dictionary to find antonyms	1	2	3	4	5
I use a dictionary to check how a word is used	1	2	3	4	5
I use a dictionary to find the origin of a word	1	2	3	4	5
I use a dictionary to help myself in translation	1	2	3	4	5
I use a dictionary to find the syntax of a word	1	2	3	4	5
I use a dictionary to find the derivatives of a word	1	2	3	4	5
I use a dictionary to find word families	1	2	3	4	5
I use a dictionary to find the meaning of an expression	1	2	3	4	5
I use a dictionary at home	1	2	3	4	5
I use a dictionary when I read a text	1	2	3	4	5

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1882 (Article)

I use a dictionary when I write a text	1	2	3	4	5
Before I buy a dictionary, I know the reason why I need it	1	2	3	4	5
Before I buy a dictionary at the bookshop, I glance through it to see what information it provides	1	2	3	4	5
I choose a dictionary because it has a lot of entries and a lot of information in each entry	1	2	3	4	5
I know what an etymological dictionary is and what it is used for	1	2	3	4	5
I know what a general dictionary is and what it is used for	1	2	3	4	5
I know what a bilingual dictionary is and what it is used for	1	2	3	4	5
I know what a dictionary of technical terms is and what it is used for	1	2	3	4	5
Before I use my new dictionary, I carefully read the introduction	1	2	3	4	5
Before I use my new dictionary, I carefully study the list of abbreviations	1	2	3	4	5
When I come across an unknown word in a text, I try to think in what form I should look it up in the dictionary	1	2	3	4	5
When I can't locate a proverb or a set phrase in the entry where I thought I would find it, I begin a new search	1	2	3	4	5
When I hear a word I don't know, I consider various spelling possibilities and look it up accordingly	1	2	3	4	5
When I can't find a word where I thought I would find it, I begin a new search until I find it	1	2	3	4	5
To see how a word is used in spoken language, I use the usage labels provided in the entry	1	2	3	4	5
When I look up a word beginning with E, I search in the first quarter pages as E is one of the first letters of the alphabet	1	2	3	4	5
When I look up a word beginning with L, I open my dictionary in the middle	1	2	3	4	5

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1882 (Article)

When I look up a word, I bear in mind its initial letter and then search where I believe this initial letter is in the dictionary	1	2	3	4	5
When I look up a word, I simply open the dictionary and see if I am near the specific initial letter	1	2	3	4	5
When I look up a word, I constantly bear it in my mind during the search	1	2	3	4	5
When I realize that the word I am looking for has various different meanings, I go through them all one by one, assisted by the example sentences	1	2	3	4	5
When I find the word that I was searching for, I return to the text to confirm that the word matches the context	1	2	3	4	5
Before I use a word I found in the dictionary when writing a text, I read all the information on the grammar of that word (conjugation, syntax) to be sure of the correct usage	1	2	3	4	5

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1883 (Article)

Making Lexicography Sustainable: Using ChatGPT and Reusing Data for Lexicographic Purposes*

Pedro A. Fuertes-Olivera, *Department of Afrikaans and Dutch, University of Stellenbosch, South Africa; International Centre for Lexicography, University of Valladolid, Spain; and Centre of Excellence in Language Technology, Ordbogen A/S, Odense, Denmark (pedro@emp.uva.es) (https://orcid.org/0000-0003-3831-5377)*

Abstract: In 2014, the International Centre for Lexicography, a research group at Valladolid signed a contract with Ordbogen A/S (a Danish language technology company) and the University of Valladolid for developing a lexicographic project, the so-called *Diccionarios Valladolid-UVa* (Fuertes-Olivera 2019, 2022a, 2022b; Fuertes-Olivera et al. 2018; Tarp and Fuertes-Olivera 2016). Each partner gave around €180,000 (the International Centre for Lexicography's contribution came from several research projects funded by the Spanish Research Agency), to be employed in the design and construction of Spanish dictionaries (in particular, a general dictionary of Spanish, a Spanish dictionary of accounting, a bilingual Spanish–English/English–Spanish dictionary and a bilingual Spanish–English/English–Spanish accounting dictionary). The above project has produced several results, with the recent publication of the *Diccionario Digital del Español* (DIDES) its most relevant result (https://diesgital.com). Within the framework of these projects, this paper offers a general introduction of the project (Section 1), refers to the concept of sustainable lexicography (Section 2), indicates that sustainability lexicography implies a better understanding of lexicographic data (Section 3), and increasing lexicographic productivity, e.g., by crafting definitions for AI translations (Section 4) and using generative AI chatbots such as ChatGPT in the day-to-day lexicographic work.

Keywords: CHATGPT, DEEPL TRANSLATE, *DICCIONARIOS VALLADOLID-UVA*, LEXICOGRAPHIC PRODUCTIVITY, SUSTAINABLE LEXICOGRAPHY, PUBLIC FUNDING, GENERATIVE AI

Opsomming: Hoe om die leksikografie volhoubaar te maak: Die gebruik van ChatGPT en die hergebruik van data vir leksikografiese doeleindes. In 2014 het die Internasionale Sentrum vir Leksikografie, 'n navorsingsgroep by Valladolid, 'n kontrak vir die ontwikkeling van 'n leksikografiese projek, die sogenaamde *Diccionarios Valladolid-UVa*, met Ordbogen A/S ('n Deense taaltegnologiemaatskappy) en die Universiteit van Valladolid onder-

---

[*] A version of this paper was presented at StellenLex 2024: Lexicography, Artificial Intelligence, Language Models and Innovation, held at the Bureau of the Woordeboek van die Afrikaanse Taal on January 24, 2024.

Lexikos 34 (AFRILEX-reeks/series 34: 2024): 123-140

teken (Fuertes-Olivera 2019, 2022a, 2022b; Fuertes-Olivera et al. 2018; Tarp en Fuertes-Olivera 2016). Elke vennoot het ongeveer €180,000 bygedra (die Internasionale Sentrum vir Leksikografie se bydrae was afkomstig van verskeie navorsingsprojekte wat deur die Spaanse Navorsingsagentskap befonds is) wat gebruik moes word in die ontwerp en samestelling van Spaanse woordeboeke (spesifiek 'n algemene Spaanse woordeboek, 'n Spaanse rekeningkundewoordeboek, 'n tweetalige Spaans–Engels/Engels–Spaanse woordeboek en 'n tweetalige Spaans–Engels/Engels–Spaanse rekeningkundewoordeboek. Bogenoemde projek het verskeie resultate tot gevolg gehad, met die onlangse publikasie van die *Diccionario Digital del Español* (DIDES) as die mees relevante produk (https://diesgital.com). Binne die raamwerk van hierdie projekte verskaf dié artikel 'n algemene inleiding tot die projek (Afdeling 1), word daar verwys na die konsep van volhoubare leksikografie (Afdeling 2), en word daar aangetoon dat volhoubare leksikografie 'n beter begrip van leksikografiese data (Afdeling 3), toenemende leksikografiese produktiwiteit, bv., deur die skep van definisies vir KI-vertalings (Afdeling 4), en die gebruik van generatiewe KI-kletsbotte soos ChatGPT in daaglikse leksikografiese take impliseer.

Sleutelwoorde: CHATGPT, DEEPL TRANSLATE, *DICCIONARIOS VALLADOLID-UVA*, LEKSIKOGRAFIESE PRODUKTIWITEIT, VOLHOUBARE LEKSIKOGRAFIE, OPENBARE BEFONDSING, GENERATIEWE KI

1. The lexicographic project *Diccionarios Valladolid-UVa*

The lexicographic project *Diccionarios Valladolid-UVa* started officially in January 2014 with the signing of a contract between the Danish language technology company Ordbogen A/S, the University of Valladolid, and the International Centre for Lexicography research group, each committing €180,000 to the project; this would be spent in the next four or five years. In the same month, we selected four part-time lexicographers, each with a 19-hour week work schedule and with an annual cost of around €25,000 (salary + labor expenses) per lexicographer. The selection process consisted of two stages, the first of which was devoted to examining the CV and English proficiency of 50 applicants. This stage resulted in the shortlisting of 10 applicants, who were given a 30-hour crash course on how to write dictionary articles and search lexicographic data with Google. These ten applicants were then asked to write 10 dictionary articles, which had been selected by the editor of the project, in a controlled environment. Their answers were then evaluated by three researchers of the International Centre for Lexicography, who selected four of the ten applicants. These four selected lexicographers started their work in March 2014; they all worked for four hours from Monday to Thursday and three hours on Friday. They were in the same room, next to the office of the editor of the project, who could check their progress and answer their queries very easily and quickly. They worked on the project until June 2020, when the Spanish Research Agency decided to stop funding the research projects they had been financing up to that time.

Cancelling public funding for the International Centre for Lexicography forced the project to change course. Since mid 2020, only the editor of the project has been engaged in it on a regular basis. The editor is totally committed to

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1883 (Article)

creating more dictionary articles for the general dictionary of Spanish, while also being open and committed to adapting the existing dictionary articles to incorporate new ideas and technological possibilities, and together with Sven Tarp, to explain the decisions taken and to explore new theoretical and practical possibilities in lexicography. It is assumed that these are truly innovative possibilities, i.e., they are the result of the development of more effective products, services, processes, technologies, and business models. Tarp (2022), for example, refers to a current project he is involved in that substantially modifies the concept of bilingual lexicography and opens the room for the use of generative AI chatbots in several lexicographic activities (see Section 4).

This article assumes that sustainable lexicography cannot be achieved without proper and regular funding and a true and effective analysis of the results obtained with the funds received. This implies a better understanding of the concept of sustainable lexicography.

2. The concept of sustainable lexicography

Sustainability in lexicography generally refers to the working conditions, re-using of lexicographic material, and financial resources that are needed for designing, making, and maintaining any lexicographic project. Kosem et al. (2021), for example, defend that semantic data should no longer exist in isolation and propose different ways for managing large, interconnected datasets. They assume that the different projects on data consolidation currently in operation will have an impact on both theoretical and practical lexicography, e.g., the outcome of the European Lexicographic Infrastructure (ELEXIS) project, which is a collaborative initiative aimed at fostering innovation and cooperation in the field of lexicography across Europe. Tiberius et al. (2024), for instance, describe the results of three international surveys that were carried out in the context of ELEXIS and that aimed at gaining insight into lexicographic practices and the lexicographers' needs in Europe.

One of the main objectives of the ELEXIS project is to integrate and make accessible the rich lexicographic resources of Europe, including dictionaries, lexical databases, and related linguistic tools and datasets. Without any doubt, such integration will reduce lexicographic costs in time and funds, and thus will make sustainable lexicography possible, especially by shifting "towards open access structured data enabling re-use and linking of dictionary data along with stand-alone lexicographic (and terminological) resources into numerous dictionary portals." (Tiberius et al. 2024: 23)

At a more down-to-earth level, lexicographic projects have to face specific drawbacks, e.g., lack of funds. Colman (2016) describes a lexicographic project (The *Algemeen Nederlands Woordenboek* (ANW), *Dictionary of Contemporary Dutch*); this project was in a similar situation to the project *Diccionarios Valladolid-UVa*; several partners initially allocated funds for the projects, but when the partners

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1883 (Article)

decided to stop funding them, they themselves had to find their own resources for continuing.

Colman (2016: 140-141) describes the ANW project as an online dictionary "through which a range of users can explore the Dutch vocabulary" and as a "linguistic data resource from which users, and especially language professionals, can extract data necessary for their research." Colman's (2016) distinction between dictionaries and lexicographic data is based on an economic and environmental interpretation of *sustainability* which demands, *inter alia*, "reuse of materials and products", "economic use of resources", "workflow optimization" and "the weighing of costs and benefits to present and future generations." The translation of the above ideas into lexicography implies that lexicographers "will need to convince funders that their investments are not a waste of time and money and that it is possible to optimize the workflow through responsible use of materials, products and financial resources" (Colman 2016: 141). In practical terms, her concept of *sustainable lexicography* implies the following:

— reusing the content of existing dictionaries, for example, adapting existing definitions to new situations;
— using links to external data, for example to a Wikipedia page;
— reusing the data of existing Dictionary Writing Systems, for example, from a monolingual dictionary to a bilingual one;
— increasing the automation of the lexicographic process itself, for example, finding "good examples" in a corpus;
— storing as much data as possible in the lexicographic database, but adapting the presentation of the data to the usage situation and user's needs (i.e., the creation of dynamic dictionary articles (Fuertes-Olivera and Bergenholtz 2011; Tarp 2011);
— making the lexicographic database usable for different purposes;
— innovating as much as possible, as shown below.

Colman (2016: 142-151) mentions four innovations in the ANW. Firstly, the traditional lexicographic definitions are complemented by a "semagram", which is basically a system of 'slots' and 'fillers' that includes all the defining characteristics of the lemma. Colman (2016: 143) claims that semagrams such as that of Table 1 (she adapts it from Moerdijk et al. 2008: 19) are useful because they enable lexicographers to make much better definitions whose additional information can "help to optimize onomasiological searches" in online dictionaries.

Secondly, the *ANW* offers lexicographic treatment of "combinatorics" and "phraseology" (Colman 2016: 144). These basically include "free combinations", "semi-fixed collocations, "fixed expressions" and "proverbs". They and the information for their lexicographic treatment is taken basically from corpora and retrieved by means of word sketches and collocation lists from the Sketch Engine; it aims at offering users "structured collocational information", i.e., "the combinations in real language use, mostly of binary combinations such as (a) noun + verb, (b) verb, verb + noun, (c) adjective + noun, and (d) adjective + *to* + verb":

this treatment will allow users, say, to find out "which verbs take *kat* (cat) as their subject and which verbs take *kat* as their object" (ibid. 145).

UPPER CATEGORY:	is an animal
CATEGORY:	is a bovine (animal)
SOUND:	moos/lows, makes a sound that we imitate with a low, long-drawn 'boe'
COLOUR:	is often black and white spotted, but also brown and white spotted, black, brown or white
SIZE:	is big
PARTS:	has an udder, horns and four stomachs: paunch, reticulum, third stomach, proper stomach
BUILD:	is big-boned, bony, large-limbed in build
FUNCTION:	produces milk and (being slaughtered) meat
PLACE:	is kept on a farm; is in the field and in winter in the byre
AGE:	is adult, has calved
PROPERTY:	is useful and tame; is considered as a friendly, lazy, slow, dumb, curious, social animal
SEX:	is female
BEHAVIOUR:	grazes and ruminates
TREATMENT:	is milked every day; is slaughtered
PRODUCT:	produces milk and meat
VALUE:	is useful

Table 1: Semagram for *cow*. Source: Colman (2016: 142)

Thirdly, the database of the ANW "functions as a kind of wordnet. For each word or word group in a particular sense, it includes related words such as hyperonyms, synonyms, antonyms, andronyms and feminines" (ibid. 146). She adds that some pragmatic information may be added, if necessary, as some research (e.g. Murphy 2013) has found that some users want more information about possible differences among synonyms, especially differences in connotation and linguistic variety. She also acknowledges that *wordnets* are difficult to process, structure and present in a dictionary. The ANW has used the thesaurus function of Sketch Engine for registering lexical and grammatical relations and includes meaning relationships "like metaphor, metonymy, generalization and specialization" when relevant (ibid. 149).

Finally, the ANW includes a large list of "simplexes", i.e., derivatives and compounds (ibid. 149), some of which are difficult to spell and some of which demonstrate the existence of regularities in word formation.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1883 (Article)

Colman (2016) mentions several drawbacks or weaknesses in each of the innovations she discusses. My view of these is mixed, as I also use some of the above ideas in the *Diccionarios Valladolid-UVa* (for example, the lemmatization of multi-word lemmas; see Fuertes-Olivera 2019 and 2022a), but I also find drawbacks that are not mentioned or assumed as such. Firstly, all the innovations discussed are language-centered, i.e., they assume that dictionaries are language artefacts and that "the art and craft of dictionary making" can be solved by offering users better and more language data. My view is that lexicographic data is much more than language data and need a proper understanding of its nature and possible functions (see Section 3). Secondly, the innovations proposed must be also analyzed in terms of lexicographic productivity, especially in terms of the money and time spent for creating the lexicographic data. For example, the application of the concept of "semagram" will be very time consuming and mostly useless as it cannot be easily implemented with many lemmas, especially with verbs, adjectives, adverbs, conjunctions, intangible nouns and so on. Instead, semagrams such as that of "cow" can be substituted by a figure and/or by using definitions from chatbots, i.e., the use of existing technology for speeding up the lexicographic process and reducing costs (see Section 4).

3. The concept of lexicographical data

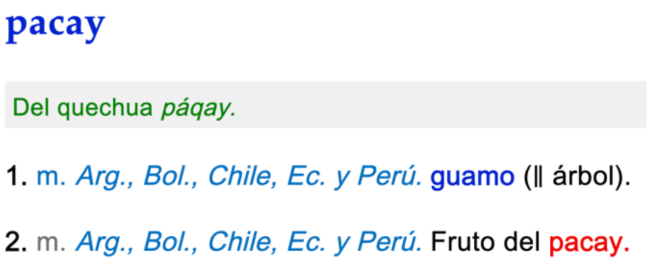
Lexicographic data are typically defined as any data that have been prepared or accepted by lexicographers and stored in a Dictionary Writing System (DWS) with the aim of helping humans and/or machines convert them into information in a *straightforward manner* (Fuertes-Olivera et al. 2018; Fuertes-Olivera and Tarp 2020). Lexicographic data can be economic resources (and hence, contribute to the concept of sustainable lexicographic) assuming that:

— They are presented in any format, e.g., as words, figures, sounds, drawings, symbols, running texts, etc.

— They may have been prepared by the lexicographers themselves or by someone else (the possibility of linking external data); this increases the "offer" of data, reduces lexicographic costs, and emphasizes that lexicographers must work with more than linguistic data.

— They must be crafted for converting them into information in a single cognitive process. This is a crucial point in our definition of lexicographic data. In these circumstances, most data in, say, existing Spanish dictionaries are not lexicographic, as they cannot be understood due to several flaws in their treatment and presentation, especially in terms of the use of a compact and traditional lexicographic style full of abbreviations, recursive definitions, and scarce relevant data (Nomdedeu-Rull and Tarp 2024).

Figure 1 shows the dictionary article **pacay** in the *Diccionario de la Lengua Española* (DLE):



pacay

Del quechua *páqay.*

1. m. *Arg., Bol., Chile, Ec. y Perú.* **guamo** (‖ árbol).

2. m. *Arg., Bol., Chile, Ec. y Perú.* Fruto del pacay.

Figure 1: The dictionary article **pacay** in the DLE

Figure 1 only informs that:

- it is a tree, whose fruit is also called "pacay";
- the tree is also called "guamo";
- it is used in two countries (Chile and Perú), and in three others identified as "Arg.", "Bol." and "Ec.";
- it derives from "quechua";
- it is "m."

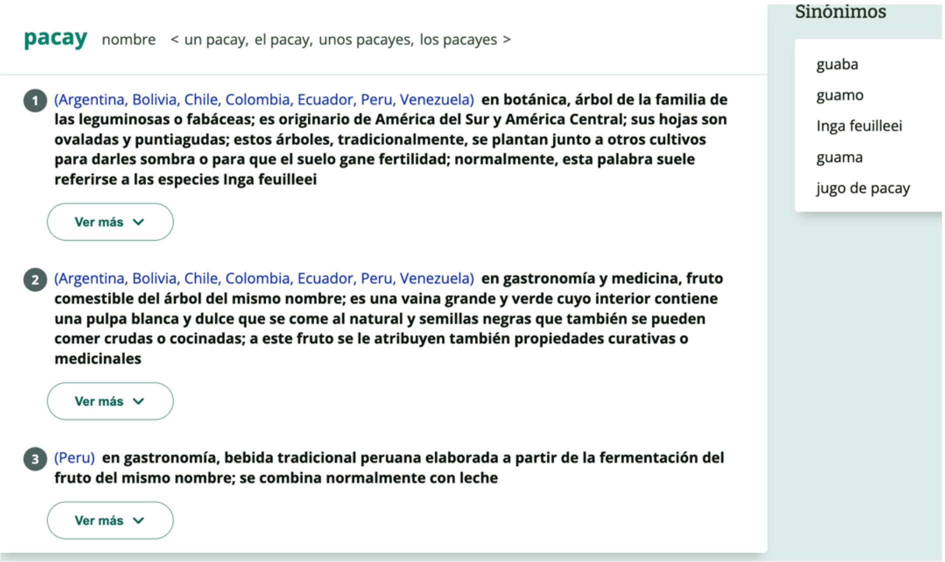
In other words, only human users who already knew what a "pacay" is can convert the data of the article into information. Such an article shows that the DLE is a "faster horse" (Tarp 2011), i.e., a printed dictionary with digital access that has not been adapted to the digital medium (Bergenholtz et al. 2009; Fuertes-Olivera 2018; Fuertes-Olivera and Bergenholtz 2011; Fuertes-Olivera and Tarp 2014; Granger and Paquot 2012). In sum, the creation of such data will make lexicography unsustainable and should rather be avoided.

DIDES uses a different approach as seen in Figure 2.

Figure 2 offers the following data about **pacay**:

- a noun with three senses — a tree, the fruit of the tree, and a traditional Peruvian drink. The tree belongs to the *Fabaceae* or *Leguminosae* family, comes from South and Central America; its leaves are oval; it is usually planted for shading other crops and fertilizing soils. The fruit is a green and big edible sheath with black seeds that can also be eaten or used in traditional medicine. The drink is typically combined with milk;
- **pacay**, and **pacayes** are its singular and plural forms;
- its accompanying articles are "un", "el", "unos", and "los";
- it has synonyms, each with its diastratic information;
- it offers examples of the three meanings used in several contexts;

- it is used in five countries: Argentina, Bolivia, Chile, Ecuador and Perú;
- it offers links to figures, e.g., to the fruit;
- it includes the buttons "ver más" (see more) or "ver menos" (see less) for accommodating the data to the size of the screen;
- it offers a difference between the tree (it is used in botany), the fruit (it is used in gastronomy and medicine), and the traditional Peruvian drink.
- It offers a complete set of clickable synonyms, which offer a complete semantic picture of the lemma, and favor cross-referencing.



pacay nombre < un pacay, el pacay, unos pacayes, los pacayes >

① (Argentina, Bolivia, Chile, Colombia, Ecuador, Peru, Venezuela) **en botánica, árbol de la familia de las leguminosas o fabáceas; es originario de América del Sur y América Central; sus hojas son ovaladas y puntiagudas; estos árboles, tradicionalmente, se plantan junto a otros cultivos para darles sombra o para que el suelo gane fertilidad; normalmente, esta palabra suele referirse a las especies Inga feuilleei**

② (Argentina, Bolivia, Chile, Colombia, Ecuador, Peru, Venezuela) **en gastronomía y medicina, fruto comestible del árbol del mismo nombre; es una vaina grande y verde cuyo interior contiene una pulpa blanca y dulce que se come al natural y semillas negras que también se pueden comer crudas o cocinadas; a este fruto se le atribuyen también propiedades curativas o medicinales**

③ (Peru) **en gastronomía, bebida tradicional peruana elaborada a partir de la fermentación del fruto del mismo nombre; se combina normalmente con leche**

Sinónimos

guaba
guamo
Inga feuilleei
guama
jugo de pacay

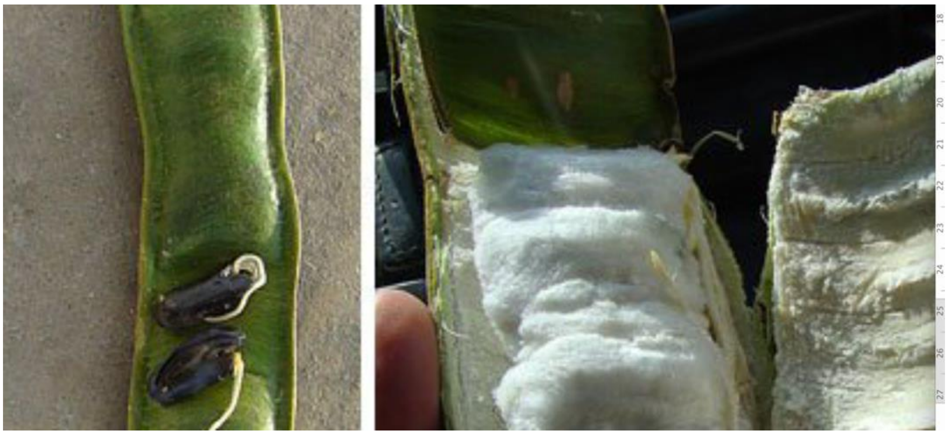


Figure 2: The lemma **pacay** in DIDES

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1883 (Article)

In sum, Figure 2 illustrates the concept of lexicographic data and its economic potential. Firstly, it can be used by humans and machines; Secondly, it really informs on meanings, forms, and functions. Thirdly, all the data can be prepared for different usages, extracted individually, and sold/licensed to third parties. Fourthly, it illustrates that lexicographic data is different from linguistic data. Finally, the use of figures, video clips and audio files, etc. may save lexicographic time and highlights the relevance of lexicographic productivity.

4. The concept of lexicographic productivity

Tarp and Fuertes-Olivera (2016) and Fuertes-Olivera (2019) have defined lexicographic productivity in economic terms; it refers to the rate at which lexicographic data are produced per unit of time, labor, or resources. For example, the lexicographic team working in the *Diccionarios Valladolid-UVa* project crafted around 35,000 dictionary articles per year for a cost of around €100,000 a year (around €3,50 per dictionary article). The number of dictionary articles crafted can be substantially increased and their cost reduced by (a) concentrating on the real nature of the lexicographic work (e.g., it is a waste of time to make a specialized dictionary without experts working on it, and (b) using existing technology for reusing existing lexicographic data and crafting new one.

The two above ideas go hand in hand and are illustrated bellow. For instance, the semagram in Table 1 (it can cost around one hour of work) can be substituted by asking ChatGPT for a definition of *cow* (below in example 1; it takes half a minute or less):

> *A cow, scientifically known as* Bos taurus, *is a domesticated mammal belonging to the family Bovidae. Cows are one of the most common types of livestock, raised globally for their ability to produce milk, meat (beef and veal), and hides, which are used in leather manufacturing. The term "cow" specifically refers to female members of the species that have given birth to at least one calf, whereas males are known as bulls, and young cattle are called calves. Cows are herbivores, primarily grazing on grasses, and they have a unique digestive system with a four-chambered stomach, which allows them to break down tough plant materials through a process called rumination.*
>
> *Cows play a significant role in agriculture and have been domesticated for thousands of years, serving not only as a source of various commodities but also as draft animals in many cultures. Beyond their economic value, cows hold profound cultural and religious significance in various societies. In some cultures, cows are revered as sacred animals and symbolize fertility, motherhood, and abundance. Their importance in human history, agriculture, and culture makes them indispensable to societies worldwide.*

Example 1: Definition of **cow** generated with ChatGPT (February 16, 2024).

Lexicographically speaking, example 1 offers a lot of information for a human lexicographer to create dictionary articles. At least, five lexicographic definitions can be crafted from the ChatGPT definition in a Spanish dictionary (the first three are in the DLE and all five in the DIDES:

1. A cow is a female herbivore mammal; it is typically found in farms and ranches throughout the world; it is raised for meat, milk, or leather, is docile and is easy to work with.
2. Cow is also a type of meat that is consumed by humans.
3. Cow is also a type of leather. This should be modified (recreated by the human lexicographer) by indicating that it is not the cow but its skin which is tanned and then used in the textile industry.
4. Cows are docile and big animals. Hence, they can be used metaphorically to refer to fat and dumb people. This meaning is informal and much used in Spanish (it is surprising that it is not included in DLE*)*.
5. Cows are cultural and religious symbols in some societies.

In addition to the above data, example 1 also indicates that its male counterpart is called a "bull", and that young animals are "calves".

Example 1 shows that the introspection and knowledge of a well-trained human lexicographer working with generative AI chatbots such as ChatGPT, Google searches (Google minitexts in Tarp and Fuertes-Olivera 2016), log files, and technology for reusing lexicographic data may enhance substantially lexicographic productivity, thus reducing costs and making the production of dictionaries cheaper. In my view, this practice is much better than working with concordances, key words, and other corpus-based or -driven technologies. Three examples illustrate this idea.

Firstly, Tarp (2022: 68), for example, is working in a project based on two experiments:

1. Using artificial intelligence to select adequate example sentences and automatically assign them to the relevant senses in a lexicographical database.
2. Using machine translation to translate L2 definitions into L1, where the translated definitions can both explain the meaning of L2 lemmata and functions as semantic differentiators when bridging from L1 to L2.

Regarding the use of Artificial Intelligence, Tarp and Henrik Hoffmann (an IT expert working at Ordbogen A/S) translated 200 Spanish definitions of the Spanish monolingual dictionary of the *Diccionarios Valladolid-UVa* project into English with the help of Google Translate and DeepL Translate (two AI-based translation tools). They found out that 78% of those translated with DeepL Translate were correct and did not need any more intervention by a lexicographer. We (Tarp, Hoffmann and myself) discussed the results and observed that automatic translations improved if the Spanish definitions were crafted adding the defining features of the *definiens* without changes of flow (e.g., without inserting non-defining relative sentences for clarifying features of the lemma being defined), using simple and clear clauses (e.g. without using subjunctives and long sentences), separating the defining features by semi-colons (instead of stops and non-defining relative clauses), and contextualizing them. Example 2 shows the legal definition of **bancarrota** (bankruptcy) in DIDES and its translation with DeepL Translate before we studied them:

- **bancarrota**
  en derecho, situación legal declarada por un juez; consiste en hacer perder a una persona, empresa, institución, organismo, etc. la disposición y administración de sus bienes, restringir su capacidad e inhabilitarle para el ejercicio de la actividad económica

- **bankruptcy**
  in law, a legal situation declared by a judge; it consists of making a person, company, institution, organization, etc. lose the disposition and administration of its assets, restricting its capacity and disqualifying it from exercising economic activity

Example 2: Initial translation of **bancarrota** with DeepL Translate

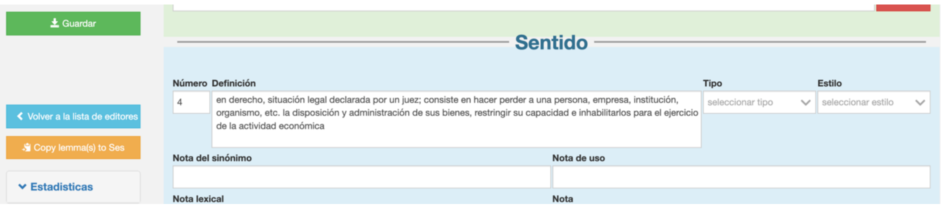
Example 2 shows that the defining features of **bancarrota** are separated by commas and semicolons. These are: (a) it is a legal term; (b) the situation occurs when a judge declares it; (c) proprietors of the asset lose it; (d) proprietors cannot continue administering the asset; (e) proprietors cannot continue with the same economic activity. The first four characteristics are perfectly translated; the fifth one, however, may be wrongly translated because the Spanish original uses "inhabilitarle" (the verb goes with a singular clitic referring a person). This can be easily corrected, e.g., by using the plural instead of the singular, as shown in example 3. The new translation is totally correct and has two interesting modifications: (a) "their" is used instead of "its" in "lose the disposition and administration of their assets" and (b) "them" is used instead of "it" in "disqualifying them":

- **bancarrota**
  en derecho, situación legal declarada por un juez; consiste en hacer perder a una persona, empresa, institución, organismo, etc. la disposición y administración de sus bienes, restringir su capacidad e inhabilitarlos para el ejercicio de la actividad económica

- **bankruptcy**
  in law, a legal situation declared by a judge; it consists of making a person, company, institution, organization, etc. lose the disposition and administration of their assets, restricting their capacity and disqualifying them from exercising economic activity

Example 3: Modified translation of **bancarrota** with DeepL Translate

Secondly, the *Diccionarios Valladolid-UVa* has also two own technologies (they were created by IT staff at Ordbogen under lexicographers' guidance) for reusing data. One of them is a "copy to" button in the DWS of the general monolingual dictionary (Figure 3; the orange button):



Número: 4

Definición: en derecho, situación legal declarada por un juez; consiste en hacer perder a una persona, empresa, institución, organismo, etc. la disposición y administración de sus bienes, restringir su capacidad e inhabilitarlos para el ejercicio de la actividad económica

Tipo: seleccionar tipo

Estilo: seleccionar estilo

Nota del sinónimo

Nota de uso

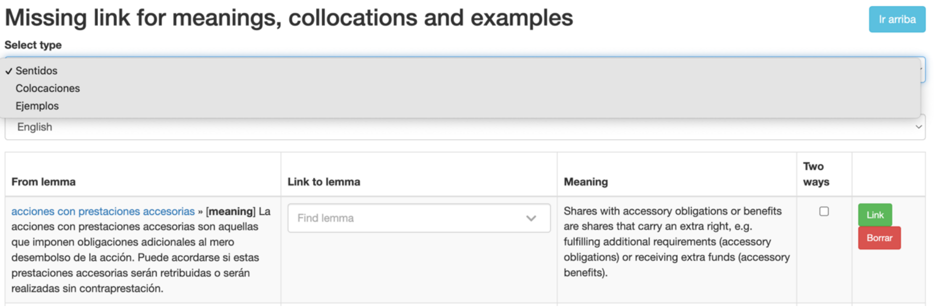
Nota lexical

Nota

Figure 3: The "copy to" button in the DWS of the Spanish general dictionary

This button allows reusing existing lexicographic data by transferring it to another DWS. The data is transferred, for example, to the Spanish side of the bilingual Spanish–English/English–Spanish dictionary of the project. It transfers the data that are correct for all types of dictionaries, typically the definition, grammar, example(s) and links of a particular meaning. It does not reuse the lemma list as each one is based on different criteria, e.g., around 80% of the lemma list of a specialized dictionary consists of multi-word lemmas, as the lemma **just-in-time inventory control system** in accounting.

The other technology in the *Diccionarios Valladolid-UVa* is a "tool" for matching lemmas with their meanings, collocations and examples in a bilingual English–Spanish/Spanish–English accounting dictionary (Figure 4):



From lemma	Link to lemma	Meaning	Two ways	
acciones con prestaciones accesorias » **[meaning]** La acciones con prestaciones accesorias son aquellas que imponen obligaciones adicionales al mero desembolso de la acción. Puede acordarse si estas prestaciones accesorias serán retribuidas o serán realizadas sin contraprestación.	Find lemma	Shares with accessory obligations or benefits are shares that carry an extra right, e.g. fulfilling additional requirements (accessory obligations) or receiving extra funds (accessory benefits).	☐	Link Borrar

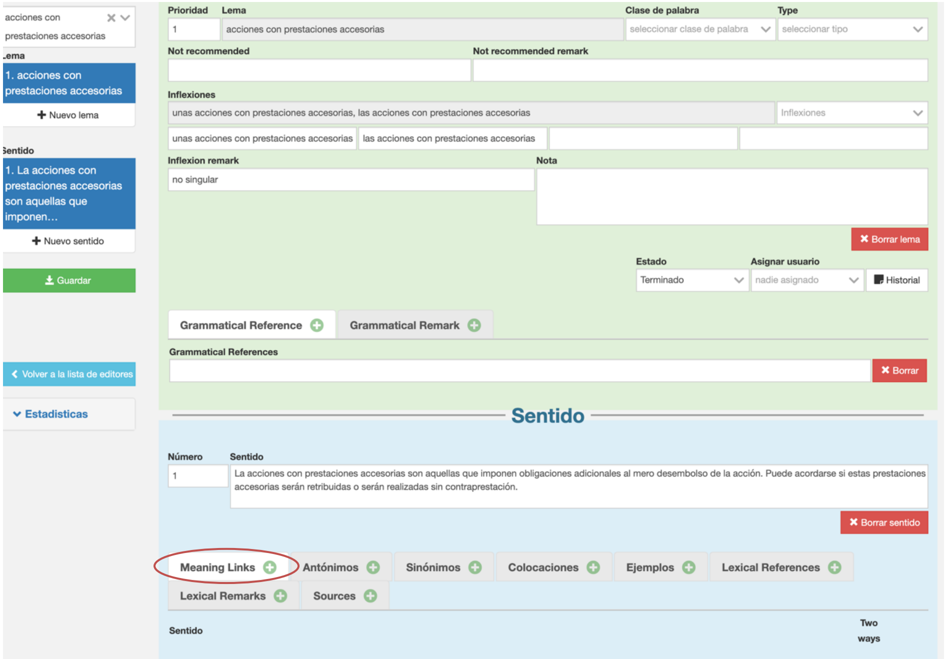
Figure 4: Tool for linking missing meanings, collocations and examples

The tool searches in a database with around 6,000 accounting lemmas, 25,000 phrases (collocations) and more than 7,000 examples in English and/or Spanish

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1883 (Article)

that are "free", i.e., they are not found as dictionary articles. They were created by a team of lexicographers and accounting experts in Denmark and Spain for a printed English–Spanish dictionary of accounting, published in 2010 (Fuertes-Olivera et al. 2010).

By clicking on the blue string of words or placing them in the search button "Link to lemma" (see Figure 4), the system automatically searches for unmatched data and, when found, creates the dictionary article (or part of it) in the language searched for and stores it in its corresponding DWS, as shown in Figure 5:



acciones con prestaciones accesorias

Lema
1. acciones con prestaciones accesorias

Sentido
1. La acciones con prestaciones accesorias son aquellas que imponen...

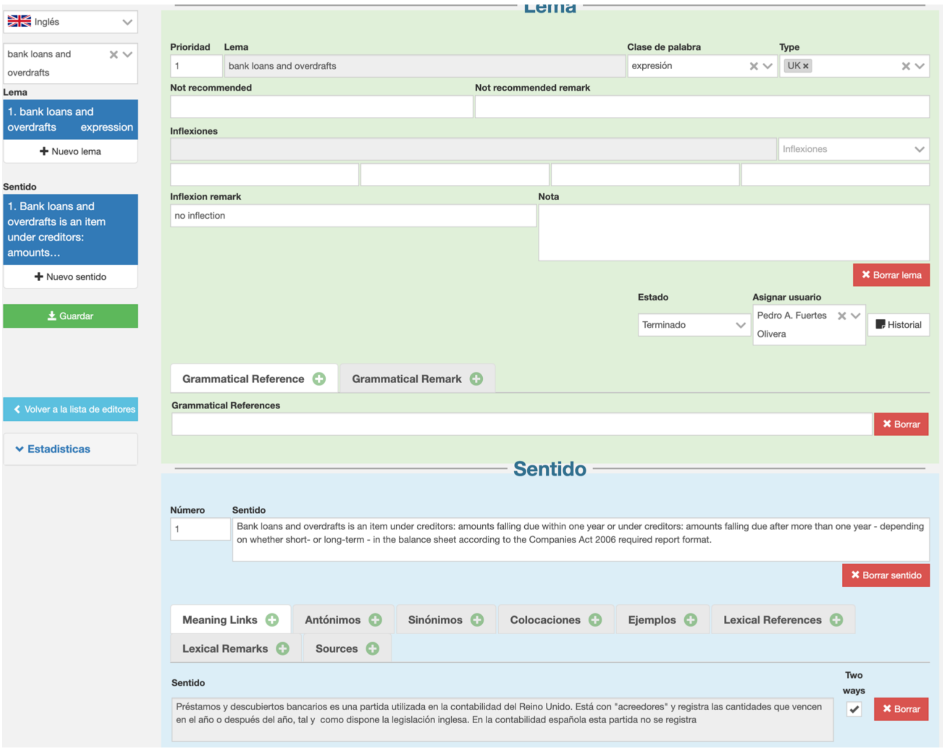
Prioridad: 1
Lema: acciones con prestaciones accesorias
Clase de palabra: seleccionar clase de palabra
Type: seleccionar tipo
Not recommended:
Not recommended remark:
Inflexiones: unas acciones con prestaciones accesorias, las acciones con prestaciones accesorias
Inflexiones: unas acciones con prestaciones accesorias | las acciones con prestaciones accesorias
Inflexion remark: no singular
Nota:
Estado: Terminado
Asignar usuario: nadie asignado
Grammatical References:

Sentido

Número: 1
Sentido: La acciones con prestaciones accesorias son aquellas que imponen obligaciones adicionales al mero desembolso de la acción. Puede acordarse si estas prestaciones accesorias serán retribuidas o serán realizadas sin contraprestación.

Figure 5: Automatic creation of the dictionary article for the accounting lemma **acciones con prestaciones accesorias** stored in the DWS of the accounting dictionary

By clicking on the button "Meaning Links" (circled in red in Figure 5), the system opens a window for writing the corresponding English lemma. If the English lemma is in the database, it will pop up and the other part of the dictionary will be automatically completed (Figure 6):



bank loans and overdrafts

1. bank loans and overdrafts expression

1. Bank loans and overdrafts is an item under creditors: amounts...

Prioridad: 1
Lema: bank loans and overdrafts
Clase de palabra: expresión
Type: UK
Inflexion remark: no inflection
Estado: Terminado
Asignar usuario: Pedro A. Fuertes Olivera

Número: 1
Sentido: Bank loans and overdrafts is an item under creditors: amounts falling due within one year or under creditors: amounts falling due after more than one year - depending on whether short- or long-term - in the balance sheet according to the Companies Act 2006 required report format.

Sentido: Préstamos y descubiertos bancarios es una partida utilizada en la contabilidad del Reino Unido. Está con "acreedores" y registra las cantidades que vencen en el año o después del año, tal y como dispone la legislación inglesa. En la contabilidad española esta partida no se registra
Two ways ☑

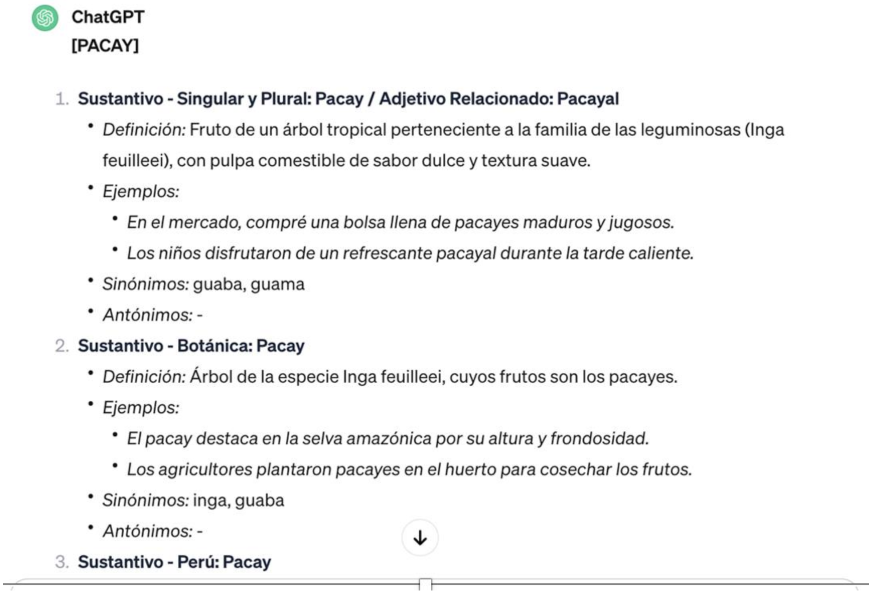
Figure 6: Automatic creation of the English lemma **bank loans and overdrafts**, with an English and Spanish definition

The two definitions always start with the lemma, followed by a verb. This makes very easy the process of searching in the tool and creating the dictionary articles automatically.

Finally, generative AI chatbots such as ChatGPT can be used for performing several lexicographic activities (see De Schryver 2023; Huete-García and Tarp 2024; Tarp and Nomdedeu-Rull 2024 for a critical analysis of the use of generative AI in lexicography). Figure 7 shows the conversation with the chatbot, initiated with an initial prompt, adapted from De Schryver (2023), about [PACAY]:

- PROMPT: Please give me lexicographic data for '[PACAY]'. 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, e.g. give me synonyms and countries in the Spanish speaking world where this word is used.

ChatGPT



ChatGPT
[PACAY]

1. **Sustantivo - Singular y Plural: Pacay / Adjetivo Relacionado: Pacayal**
   - *Definición:* Fruto de un árbol tropical perteneciente a la familia de las leguminosas (Inga feuilleei), con pulpa comestible de sabor dulce y textura suave.
   - *Ejemplos:*
     - *En el mercado, compré una bolsa llena de pacayes maduros y jugosos.*
     - *Los niños disfrutaron de un refrescante pacayal durante la tarde caliente.*
   - *Sinónimos:* guaba, guama
   - *Antónimos:* -
2. **Sustantivo - Botánica: Pacay**
   - *Definición:* Árbol de la especie Inga feuilleei, cuyos frutos son los pacayes.
   - *Ejemplos:*
     - *El pacay destaca en la selva amazónica por su altura y frondosidad.*
     - *Los agricultores plantaron pacayes en el huerto para cosechar los frutos.*
   - *Sinónimos:* inga, guaba
   - *Antónimos:* -
3. **Sustantivo - Perú: Pacay**

PF
PROMPT: Please tell me in which American countries this word is used because, to the best of my knowledge, PACAY is not used in Spain.

ChatGPT:

- La palabra "pacay" y sus significados son más comunes en países de América Latina, especialmente en regiones tropicales y subtropicales. Aunque las acepciones específicas pueden variar, la palabra es generalmente reconocida en contextos relacionados con frutas tropicales, árboles y bebidas tradicionales. Aquí hay algunos países de América donde se utiliza la palabra "pacay":

1. Perú: En Perú, además de referirse al fruto y al árbol, "pacay" se utiliza específicamente para denominar una bebida tradicional fermentada hecha a partir del jugo de pacay.
2. Ecuador: La palabra "pacay" es comúnmente utilizada en Ecuador para referirse al fruto y al árbol.
3. Colombia: En algunas regiones de Colombia, como en la Amazonía, es probable encontrar el uso de la palabra "pacay" para hablar de esta fruta tropical.
4. Venezuela: La palabra también se utiliza en Venezuela para referirse al árbol y al fruto.

- Es cierto que la palabra "pacay" y sus acepciones no son comunes en España, donde el léxico y las costumbres relacionadas con frutas tropicales pueden diferir. La variación en el uso de palabras es común en español debido a las diferencias culturales y geográficas entre las distintas regiones hispanohablantes.

Figure 7: Dialogue with ChatGPT on the Spanish word **pacay**

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1883 (Article)

The above dialogue shows that working with chatbots such as ChatGPT has advantages and disadvantages. The former is that it can increase productivity, reduce lexicographic costs in both time and money, and allow searching for data that can be difficult to obtain, e.g., a particular meaning of a word which may be only used in one of the countries where Spanish is spoken (Spanish is spoken by more than 500 million native speakers). In future, it will be necessary to refine the practice of making prompts asking the chatbot for such data. Disadvantages are also well-known (see De Schryver 2023, Rundell 2023 and Huete-Garcia and Tarp 2024): hallucinations may be widespread and therefore, it is advisable to double check the data obtained with a chatbot before using them. These potential disadvantages, however, cannot make us forget the usefulness of Chatbots, e.g., example 2 was crafted with data shown in Figure 7.

5. Conclusion

This article has discussed the concept of sustainable lexicography in a rather different fashion to the one initially published by Colman (2016). The approach used here attempted to show 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. The use of chatbots and other AI functionalities merit our consideration. I have no doubt that these will improve in time and that some of the qualms expressed these days by well-known scholars such as Vossen, (2022), Chomsky et al. (2023), McKean and Fitzgerald (2023) and Rundell (2023) will fade away.

Acknowledgment

Thanks are due to the participants in the symposium StellenLex 2024, held at Bureau of the Woordeboek van die Afrikaanse Taal, specially to the organizers Professors Rufus H. Gouws, Theo D. Bothma and Sven Tarp. I also thank the reviewers of this article and the editor of *Lexikos*, André du Plessis, for their comments on a previous draft of this paper.

Bibliography

Bergenholtz, H., S. Nielsen and S. Tarp (Eds.). 2009. *Lexicography at a Crossroads.* Bern: Peter Lang.

ChatGPT: https://chat.openai.com/ (Access: February, 2024)

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 (Access: April, 2024)

Colman, L. 2016. Sustainable Lexicography: Where to Go from Here with the *ANW* (*Algemeen Nederlands Woordenboek*, an Online General Language Dictionary of Centemporary Dutch? *International Journal of Lexicography* 29(2): 139-155.
https://doi.org/10.1093/ijl/ecw008

DeepL Translate: https://www.deepl.com/es/translator (Access: April, 2024)

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

DIDES. *Diccionario Digital del Español*. https://diesgital.com/ (Access: April, 2024)

DLE. *Diccionario de la Lengua Española*. RAE. https://dle.rae.es/ (Access: April, 2024)

Fuertes-Olivera, Pedro A. (Ed.). 2018. *The Routledge Handbook of Lexicography.* London/New York: Routledge.
https://doi.org/10.4324/9781315104942

Fuertes-Olivera, Pedro A. 2019. Designing and Making Commercially Driven Integrated Dictionary Portals: The *Diccionarios Valladolid-UVa. Lexicography* 6: 21-41.
https://doi.org/10.1007/s40607-019-00056-8

Fuertes-Olivera, Pedro A. 2022a. The Mental Lexicon in Lexicography: *The Diccionarios Valladolid-UVa. Lexikos* 32(1): 118-140.
DOI: https://doi.org/10.5788/32-1-1712

Fuertes-Olivera, Pedro A. 2022b. Theoretical, Technological and Financial Challenges: Some Reflections for Making Online Dictionaries. Jackson, Howard (Ed.). 2022. *The Bloomsbury Handbook of Lexicography:* 361-374. London/New Delhi/New York/Sydney: Bloomsbury Academic.
10.5040/9781350181731.ch-021

Fuertes-Olivera, Pedro A. and H. Bergenholtz (Eds.). 2011. *e-Lexicography*: *The Internet, Digital Initiatives and Lexicography*. London/New York: Continuum.
10.5040/9781474211833

Fuertes-Olivera, Pedro A. and S. Tarp. 2014. *Theory and Practice of Specialised Online Dictionaries. Lexicography versus Terminography*. Berlin/Boston: De Gruyter.
https://doi.org/10.1515/9783110349023

Fuertes-Olivera, Pedro A. and S. Tarp. 2020. A Window to the Future: Proposal for a Lexicographically-assisted Writing Assistant. *Lexicographica* 36: 257-286.
https://doi.org/10.1515/lex-2020-0014

Fuertes-Olivera, Pedro A., S. Tarp and P. Sepstrup. 2018. New Insights in the Design and Compilation of Digital Bilingual Lexicographical Products: The Case of the *Diccionarios Valladolid-UVa*. *Lexikos* 28: 152-176.
https://doi.org/10.5788/28-1-1460

Fuertes Olivera, Pedro A., P. Gordo Gómez, M. Niño Amo, A. de los Ríos Rodicio, A. Sastre Ruano, S. Tarp, M. Velasco Sacristán, S. Nielsen, L. Mourier and H. Bergenholtz. 2010. *Diccionario de Contabilidad Inglés–Español*. Navarra: Thomson Reuters-Aranzadi.

Granger, S. and M. Paquot (Eds.). 2012. *Electronic Lexicography.* Oxford: OUP.
https://doi.org/10.1093/acprof:oso/9780199654864.001.0001

Huete-García, A. and S. Tarp. 2024. Training AI-based Writing Assistant for Spanish Learners: The Usefulness of Chatbots and the Indispensability of Human-assisted Intelligence. *Lexikos* 34(1): 21-40.
https://doi.org/10.5788/34-1-1862

Kosem, I., S. Krek and P. Gantar. 2021. Semantic Data Should no Longer Exist in Isolation: The Digital Dictionary Database of Slovenian. *Euralex 2020. Lexicography for Inclusion, 7–11 September 2021, Virtual.*
https://elex.is/euralex2020/ (Access: April, 2024)

McKean, E. and W. Fitzgerald. 2023. The ROI of AI in Lexicography. *Proceedings of the 16th International Conference of the Asian Association for Lexicography: Lexicography (Asialex 2023 Proceedings), 22–24 June 2023, Seoul, Korea: Artificial Intelligence, and Dictionary Users:* 10-20. Seoul: Yonsei University.

Moerdijk, F., C. Tiberius and J. Niestadt. 2008. Accessing the ANW Dictionary. Michael Zock and Chu-Ren Huang (Eds.). 2008. *Coling 2008: Proceedings of the Workshop on Cognitive Aspects of the Lexicon (COGALEX 2008),* Manchester, 24 August 2008: 18-24. Manchester, UK: Coling 2008 Organizing Committee.
https://aclanthology.org/W08-1900/

Murphy, M.L. 2013. What We Talk about When We Talk about Synonyms (And What It Can Tell Us about Thesauruses). *International Journal of Lexicography* 26(3): 279-304.

Nomdedeu-Rull, A. and S. Tarp. 2024. *Introducción a la lexicografía en español. Funciones y aplicaciones.* London: Routledge.

Rundell, M. 2023. Automating the Creation of Dictionaries: Are We Nearly There? *Proceedings of the 16th International Conference of the Asian Association for Lexicography: Lexicography (Asialex 2023 Proceedings), 22–24 June 2023, Seoul, Korea: Artificial Intelligence, and Dictionary Users:* 1-9. Seoul: Yonsei University.

Tarp, S. 2011. Lexicographic and Other e-Tools for Consultation Purposes: Towards the Individualization of Needs Satisfaction. Fuertes-Olivera, Pedro A. and Henning Bergenholtz (Eds.). 2011. *e-Lexicography: The Internet, Digital Initiatives and Lexicography:* 54-70. London/New York: Continuum.

Tarp, S. 2022. Turning Bilingual Lexicography Upside Down: Improving Quality and Productivity with New Methods and Technology. *Lexikos* 32: 66-87.
https://doi.org/10.5788/32-1-1686

Tarp, S. and Pedro A. Fuertes-Olivera. 2016. Advantages and Disadvantages in the Use of Internet as a Corpus: The Case of the Online Dictionaries of Spanish Valladolid-UVa. *Lexikos* 26: 273-295.
https://doi.org/10.5788/26-1-1349

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* 97: 309-321.
https://dx.doi.org/10.5209/clac.91985

Tiberius, C., J. Kallas, S. Koeva, M. Langemets and I. Kosem. 2024. A Lexicographic Practice Map of Europe. *International Journal of Lexicography* 37(1): 1-28.
https://doi.org/10.1093/ijl/ecad023

Vossen, P. 2022. ChatGPT Is a Waste of Time. *VU Magazine,* 22 December 2022.
https://vumagazine.nl/professor-piek-vossen-chatgpt-is-a-waste-of-time?lang=e

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1901 (Article)

This, Thing, Fervor, Fulfilment: The Treatment of Pronunciation and Spelling in Dictionaries of the Slovenian Immigration

Donna M.T.Cr. Farina, *Department of Multicultural Leadership, New Jersey City University, New Jersey, USA*
(dfarina@njcu.edu) (https://orcid.org/0000-0001-5695-7782)

Marjeta Vrbinc, *Faculty of Arts, University of Ljubljana, Slovenia*
(marjeta.vrbinc@ff.uni-lj.si) (https://orcid.org/0000-0002-6866-6023)
and
Alenka Vrbinc, *School of Economics and Business, University of Ljubljana, Slovenia*
(alenka.vrbinc@ef.uni-lj.si) (https://orcid.org/0000-0002-7330-4158)

Abstract: In the second half of the 19th century, dictionaries increased in importance among Americans. They began to be perceived as authorities by the U.S. population; users expected them to provide answers to their questions about language. At the turn of the 19th century into the 20th, on both sides of the Atlantic, the first independent Slovenian publications appeared, intended for Slovenian immigrants to the U.S. The goal of the present article is to examine the treatment of pronunciation and spelling, both in the front matter and in the body of dictionaries of the Slovenian immigration. We examine four dictionaries created by three authors (Kubelka 1904, Kubelka 1912b, Košutnik 1912, Kern 1919). They were published at a time when there were no readily available resources on English pronunciation or spelling written in Slovenian. This article documents the dictionary authors' explanations of pronunciation and how these explanations were presented to the intended audience. It also documents the treatment of spelling of words with predominantly American and predominantly British variants, at a time when both variants were widely circulating within American society.

Keywords: SLOVENIAN IMMIGRATION TO U.S., BILINGUAL DICTIONARIES, ENGLISH–SLOVENIAN DICTIONARIES, SLOVENIAN–ENGLISH DICTIONARIES, PRONUNCIATION, IPA, RESPELLING, SPELLING, BRITISH ENGLISH, AMERICAN ENGLISH

Opsomming: ***"This", "thing", "fervor", "fulfilment":*** **Die hantering van uitspraak en spelling in Sloweense immigrasiewoordeboeke.** In die tweede helfte van die 19de eeu het woordeboeke vir Amerikaners belangriker geword. Die Amerikaanse bevolking het woordeboeke as gesaghebbend begin beskou; gebruikers het verwag dat hulle antwoorde op hul taalvrae sou verskaf. Teen die draai van die 19de eeu het die eerste onafhanklike Sloweense publikasies, bedoel vir Sloweense immigrante na die V.S.A., aan beide kante van die Atlantiese

Oseaan verskyn. Die doel van hierdie artikel is om die hantering van uitspraak en spelling, beide in die voorwerk en in die sentrale deel van Sloweense immigrasiewoordeboeke, te bestudeer. Ons ondersoek vier woordeboeke (Kubelka 1904, Kubelka 1912b, Košutnik 1912, Kern 1919) wat deur drie outeurs saamgestel is. Hulle is in 'n tydperk gepubliseer toe hulpmiddels vir Engelse uitspraak of spelling nie geredelik in Sloweens beskikbaar was nie. In hierdie artikel word die woordeboekouteurs se toeligting rakende uitspraak asook die aanbieding van hierdie toeligting vir die teikengebruikers, gedokumenteer. Die hantering van die spelling van woorde met oorwegend Amerikaanse en oorwegend Britse variante, op 'n tydstip toe beide variante wydverspreid in die Amerikaanse samelewing voorgekom het, word ook gedokumenteer.

Sleutelwoorde: SLOWEENSE IMMIGRASIE NA DIE V.S.A., TWEETALIGE WOORDEBOEKE, ENGELS–SLOWEENSE WOORDEBOEKE, SLOWEENS–ENGELSE WOORDEBOEKE, UITSPRAAK, IFA, HERSPELLING, SPELLING, BRITSE ENGELS, AMERIKAANSE ENGELS

1. Introduction

In the second half of the 19th century, dictionaries increased in importance among Americans for various reasons: the growth of popular linguistic knowledge; industrialization and the growth of technology, which resulted in new vocabulary; and population growth and the expansion of public education as a means of self-improvement, which resulted in a huge demand for books which taught immigrants and others how to speak and write correctly (Landau 2001: 85; Shapiro 2020). Adams (2015: 25) calls this developing attitude of the American public as "linguistic insecurity", an attitude that "was pervasive when Webster published his big dictionary of American English in 1828, urgent as America welcomed waves of immigrants in the 19th and earlier 20th centuries, and relevant to the present day". In short, dictionaries began to be perceived by the U.S. population as authorities, since users expected them to provide answers to their questions about usage, pronunciation, etymology, etc.; in other words, users expected dictionaries to be prescriptive (Landau 2001: 85).

Stanonik (1996) recounts what was happening on both sides of the Atlantic, moving toward the turn of the 19th century into the 20th. The first independent Slovenian publications appeared, intended for Slovenian immigrants to the U.S. Their aim was to provide basic useful information on the English language, the U.S. constitution and organization of the government, the American monetary system and economy. Some of these were printed in America, others in Slovenia. In 1879, the *Slovensko–angleška slovnica* [Slovenian–English Grammar] came out in Tower, Minnesota. Its author was Peter Jeram, a priest in Wabasha, Minnesota. This is most likely the first book in Slovenian that was printed in America (first reprint 1895). This was followed by an anonymous work entitled *Angleščina brez učitelja. Pomočna knjiga za izseljence* [English without a Teacher, a Handbook for Emigrants] (Ljubljana, 1895);[1] another work published in Ljubljana (1904, second edition 1912) is Silvester Košutnik's *Ročni slovensko–angleški in angleško-slovenski slovar* [A Pocket Slovenian–English and English–Slovenian Dictionary]. These reference works were followed by similar

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1901 (Article)

books by Viktor Kubelka (two dictionaries and a phrasebook), Frank Javh-Kern (a dictionary; Cleveland 1919), Kazimir Zakrajšek (a children's alphabet book; New York 1917, Chicago 1923) and Ivan Mulaček (a quasi-textbook, a self-study grammar; Ljubljana 1930). Slovenian speakers, insecure about their English or desiring to learn it, could use these dictionaries and grammar books (and possibly other books unknown to us) published before World War II. The Slovenian–American users were usually not in a position to evaluate the quality of the books; most had only primary education, so would not be aware of lapses or inconsistencies.

2. Aim of the study

Many authors (Landau 2001; Béjoint 2010; Adams 2015) have explored, in the period from the early- or mid-1800s into the 20th century, the relationship of the American public with the dictionaries being created for it. Most or all of this exploration appears to be focused solely on monolingual speakers of American English (AmE) and their use of the monolingual English dictionaries available to them. However, the period 1860–1915 is also a time of a mass migration to the U.S. The Library of Congress (n.d.) notes that "Between 1900 and 1915, more than 15 million immigrants arrived in the United States. That was about equal to the number of immigrants who had arrived in the previous 40 years combined". These immigrants were mostly from Europe, including Slovenians who are our focus here. This is an astounding figure, since the 1860 census estimates just under 31,450,000 people living in the U.S. (United States Census Bureau, n.d.). It is estimated that around 86,000 Slovenians from the area of Carniola came to the U.S. from 1892 to 1913.

While certainly the newcomers had priorities (such as eating and finding work) that took precedence over mastery of the English language, nevertheless they were highly motivated to learn at least survival English in order to function in the new American society themselves and to gain advantages (or avoid disadvantages, including prejudice) for their children. In their daily lives back in the territories of what would become modern Slovenia, these immigrants usually were not extensively educated or used to relying on books. So for them, the sparse bilingual resources they had at their disposal, in the form of sometimes amateurish grammar books (e.g., Jeram 1895), phrasebooks (e.g., Kubelka 1912a), and dictionaries, were imbued with the same high level of authority that monolingual Americans gave their large English dictionaries and other language resources.

Bilingual reference books with English from this period in American history, whether for the Slovenian immigration or for one of the many other immigrant groups, have not received much attention from historians of lexicography or others. While the books may appear elementary, unprofessional, or even primitive to the modern eye, they have much to tell us about American attitudes toward education, reading, and what constitutes correct speech. While the present article addresses just a few of these books — four dictionaries for the Slovenian immigrant community — it should be considered a necessary small step in helping to fill our knowledge gap concerning what new bilingual Americans thought about language and dictionaries.

The body of a dictionary contains a list of headwords or lemmata (i.e., one element of the macrostructure). Each headword is accompanied by multiple pieces of information, which together with the headword constitute the dictionary entry. The microstructure is the internal organization of the various pieces of information which are contained in the dictionary entry. The microstructure consists of detailed information about the headword, with comments on its formal properties (spelling, pronunciation and grammar) and its semantic properties (definition, usage and etymology), as well as other information. The goal of the present article is to examine the treatment of pronunciation and spelling, both in the front matter as well as in the body of the dictionaries of the Slovenian immigration. This, we consider, is in line with the aim of further elucidating attitudes on the correct use of English in speech and writing. We envisage that future work will address other aspects of microstructure, namely meaning and equivalence in the dictionaries intended for new Americans of Slovenian origin.

3. Slovenian and English bilingual dictionaries

The present article examines four of the dictionaries produced at the height of the Slovenian immigration to the U.S.[2] The dictionaries, by the authors Victor/ Viktor J. Kubelka (1904, 1912b), Silvester Košutnik (1912) and Fran/Frank J. Kern (1919), were created with this target population of Slovenian newcomers in mind. One of these authors, Kubelka, emphasizes in his (Slovenian-language) preface (1912b) the great demand for all of his various Slovenian–English books. In the (Slovenian-language) preface to the (second) 1912 edition of his dictionary, Košutnik says: "Finally, I need to mention — and this should be considered a proof — how welcome this dictionary was for our emigrants: the first edition was sold out in a short time; due to great demand, I was forced to prepare a revised edition" (p. 4). Kubelka and Kern were Slovenian immigrants and published in the U.S.; Košutnik, on the other hand, published in Ljubljana, Slovenia and never emigrated to America.

3.1 Dictionary sources

The lexicographic sources used by these authors in the compilation of their bilingual dictionaries are a complete mystery; it is not even clear that there are sources for some of the dictionaries being examined. For these authors of varying degrees of education or knowledge of English and/or Slovenian, it was most likely difficult or impossible to obtain dictionaries that might be relevant to their lexicographic tasks. While it is probable that all three authors knew

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1901 (Article)

German well, it is not clear that they were familiar with the German lexicographic tradition or that they had access, particularly in the American context, to English and German bilingual dictionaries. What is more, in the first two decades of the 20th century, the study of English was not widespread in Europe and the number of possible English and German resources was not large.[3] There are only two instances where, in the dictionaries being inspected here, an author mentions any other dictionary at all: Kern (1919, English preface, p. III) mentions the *New Standard Dictionary* (first published in 1913) as an influence on his treatment of pronunciation and spelling. Furthermore, he mentions (1919, Slovenian preface, p. V) that, while his own dictionary is appropriate for learners, people with a greater knowledge of English should rely on the *New Standard Dictionary* or on "Webster".

In our discussions of the dictionaries below, page numbers are often included with examples. We chose this practice for two reasons. First, having a page number available facilitates manual searches when a digital manuscript does not have search-and-find capacities. Second, in Kubelka (1904), the alphabetical order of entries is not always followed and page number information is helpful for expediting searches.

3.1.1 Kubelka's 1904 and 1912 dictionaries

Kubelka's first attempt at a reference book is his *Slovensko–angleški žepni rečnik v olajšavo naučenja obeh jezikov, Slovenian–English Pocket Dictionary to Facilitate the Study of Both Languages* (1904). He would go on to publish a phrasebook[4] for use by new arrivals. His final and most comprehensive work is *Slovensko–angleška Slovnica, Tolmač, Spisovnik in Navodilo za Naturalizacijo, Angleško-Slovenski in Slovensko–Angleški Slovar, Slovenian–English Grammar Interpreter, Letterwriter and Information on Naturalization, English–Slovenian and Slovenian–English Dictionary* (1912b), that included a much larger bidirectional bilingual dictionary as well as other components for language learning.

The central part of Kubelka (1904) is the Slovenian–English dictionary itself (pp. 24-122; 99 pages), in which Slovenian lemmata (column 1) are followed by English equivalents (column 2) and the pronunciation of the English equivalent (column 3). Only the pronunciation is in italics; the lemmata and equivalents are in normal typeface. The larger Kubelka (1912b) has both Slovenian–English and English–Slovenian dictionary components, in addition to other parts. The English–Slovenian dictionary appears first (pp. 210-295; 86 pages) and the Slovenian–English dictionary (pp. 296-423; 128 pages) is second. Unlike Kubelka (1904), Kubelka's new Slovenian–English dictionary (1912b) is designed more professionally. Instead of the three parallel columns of the 1904 work, we now have actual dictionary entries with components in the following order: Slovenian lemma, English equivalent, and English pronunciation of the equivalents in parentheses. Note that we would not expect any modern bilingual dictionary intended for encoding (as the Slovenian–English sections both in

Kubelka (1904) and (1912b) surely were intended) to list the pronunciation of equivalents. Normally, pronunciation immediately follows the lemma, so experienced users would not expect pronunciation information about English equivalents. However, this move of Kubelka's to include pronunciation of equivalents is quite reasonable (albeit unorthodox), since his target immigrant population did not consist of experienced dictionary users. His audience would sorely need such guidance on American pronunciation. Finally, we can point out that Kubelka's (1912b) Slovenian–English section uses boldface for the lemmata, followed by normal typography for equivalents and pronunciation.

The English–Slovenian section of Kubelka (1912b) is more basic in design than its Slovenian–English section and uses typography differently. It contains English lemmata in normal typography (no boldface) followed by one or more Slovenian equivalents (usually no more than three) in italics. There is no pronunciation given in the English–Slovenian section, either of the English lemmata or of the equivalents. While a Slovenian speaker does not need pronunciation for the Slovenian equivalents, it would have been quite helpful for the English lemmata. In a modern English–Slovenian bilingual dictionary, certainly pronunciation of English lemmata would be necessary and included.

3.1.2 Košutnik's 1912 dictionary

Košutnik's *Ročni slovensko–angleški in angleško slovenski slovar: Zlasti namenjen izseljencem v Ameriko* [A Pocket Slovenian–English and English–Slovenian Dictionary: Intended for Immigrants to America] was first published in 1904; the second edition appeared in 1912. Stanonik (1996) called the 1904 dictionary "anonymous", an understandable error given that Košutnik often used his initials only, making it appear as if the 1904 book was written by an anonymous author. Both the 1904 and 1912 books are listed in the Slovenian National University Library catalogue, with the 1904 edition listed as authored by Košutnik. The titles of the 1904 and 1912 books are identical and the national catalogue calls the 1912 edition a reprint. The present analysis is based on the 1912 version.

Košutnik's dictionary is bidirectional, first Slovenian–English (pp. 5-63; 59 pages) and then English–Slovenian (pp. 64-148; 85 pages). The Slovenian–English dictionary has a simple microstructure: lemma in Slovenian, dash, equivalent(s) in English, period; there is no use of either boldface or italics. The English–Slovenian dictionary structure is English lemma, comma, followed by a simplified pronunciation (with primary stress noted as a superscript dash after the stressed syllable), then a dash and the Slovenian equivalent(s). Essentially the structure of both sections is the same, only the English–Slovenian section adds on pronunciation of the lemmata.

3.1.3 Kern's 1919 dictionary

Kern's *A Complete Pronouncing Dictionary of the English and Slovene Languages for*

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1901 (Article)

General Use, Popoln angleško-slovenski besednjak z angleško izgovarjavo is the only one of the four dictionaries that has the Slovenian title second and the English title first. Košutnik did not include an English title at all and Kubelka's two dictionaries list the Slovenian title followed by the English title. We presume that by this later publication date (seven years after the 1912 publication of Kubelka and of Košutnik), the Slovenian community (and Kern himself) had transitioned further on the path to English dominance and that this was reflected in the title choice.

Kern's work consists solely of a monodirectional English–Slovenian dictionary (pp. 1-270). The earliest dictionary, Kubelka (1904), was monodirectional with only Slovenian–English; the middle dictionaries of 1912, Košutnik's and Kubelka's, were bidirectional, and Kern (1919) is only English–Slovenian. Like with the selection of titles, this arguably indicates the movement of the Slovenian community over to English.

At 270 pages, Kern's dictionary is more than twice the size of the largest (monodirectional) component in another dictionary; the next largest is Kubelka's Slovenian–English component with 128 pages. His dictionary is more comprehensive and professional than its predecessors. The design is the most "dictionary-like" of the four examined here; it resembles most what one would expect from later 20th-century print dictionaries. The English lemma is printed in bold and followed by a comma, pronunciation in parentheses, another comma and then the equivalent(s). The other dictionary that uses boldface in lemmata is Kubelka (1912b) — and only for Slovenian lemmata in the Slovenian–English section.

4. Pronunciation

Pronunciation can be defined as the form, production, and representation of speech. It is the phonological counterpart of spelling (orthography); i.e., its shape in the medium of sound in contrast with its shape in the medium of writing (Jackson 2002). While pronunciation and spelling may be counterparts, they are also inextricably linked in complex ways. They both present endless problems and difficulties for the language learner and are a topic of discussion in all of the Slovenian and English bilingual dictionaries targeted here. Their treatment varies from one dictionary to another; certainly considerations of the authors' education and experience inform the divergent treatments, as we will address later in the final discussion and conclusions.

4.1 IPA versus respellings

All of the dictionaries under consideration provide their own respelling systems for indicating pronunciation. It is not surprising that none of them use the International Phonetic Alphabet (IPA) to represent English, given that it was only developed in the late 19th century (Jackson 2002) and not adapted for

English until 1904 (Passy 1904); later the International Phonetic Association published samples in many languages, including AmE ("The Principles of the International Phonetic Association" 1912). When Kubelka created his first dictionary, IPA had barely been established; by 1912 and 1919 when the other three dictionaries examined here were published, the alphabet was still hardly used in America and had yet to be adapted for Slovenian.

4.2 Advice is cheap

In addition to their instructions on how to render one or another sound in English, sometimes the authors provide platitudes designed to motivate the learner in persisting with the difficulties of English pronunciation. Kubelka (1904) refrains from this type of advice entirely, but Kubelka (1912b) writes: "English pronunciation is difficult; correct pronunciation is possible only with persistent practice" (Slovenian-language preface, p. 11). Kubelka admits that his pronunciation advice is not perfect, but adds that if he wanted to present the pronunciation of individual phonemes, especially vowels, he would have had to use many symbols, which would make learning English difficult and cause people not to use his book. As we will see, the lengthy pronunciation section of Kubelka (1912b) could have dissuaded potential users from tackling the English language.

Košutnik (1912) is at the opposite extreme as Kubelka (1912b), and concludes his very brief pronunciation remarks by saying that, while he could bring in many more examples, they would not be very useful for "those layers of society for which this booklet is intended" (Slovenian-language preface, p. 4). Such people, he says, should "listen carefully" when someone speaks English and that they "should use any opportunity offered to them to practice pronunciation" (p. 4).

Kern (1919) is perhaps the most circumspect of our three authors. In his Slovenian preface, he modestly warns the reader not to take his pronunciation instructions too literally: "The diacritics and pronunciation given are only for First Aid". He adds: "It is necessary to learn by means of listening and speaking" (p. IV).

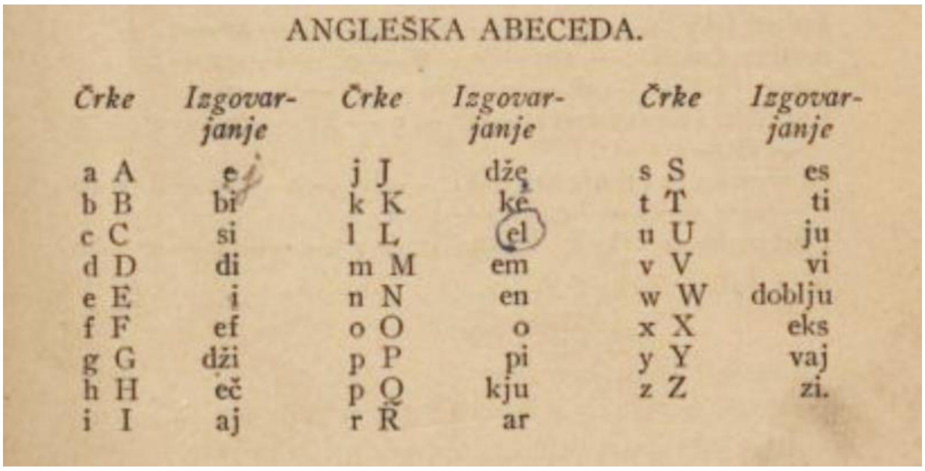
4.3 Kubelka 1904 and 1912b

In both his English- and Slovenian-language prefaces, Kubelka (1904) discusses only the pronunciation of /ð/ and explains that, in order to avoid presenting a too-complex pronunciation rule system, he renders it as *dz*. Kubelka (1912b) has only a Slovenian preface; in it, /ð/ is not mentioned. However, in the front matter as part of a lengthy four-page treatment of consonants, Kubelka (1912b) has a revised rendering of this sound as either *d* or *t*, and gives as an example *father* /fadr/ (p. 26).

Following the two prefaces, in the Slovenian-language front matter to his

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1901 (Article)

Slovenian–English dictionary, Kubelka (1904) provides each letter of the English alphabet followed by a spelling pronunciation (p. 11). For example, the letter "A" is given the pronunciation *ej*, which would be the normal rendering of this diphthong in any Slovenian dictionary prior to the use of IPA. Less orthodox renderings are "H" *eč*, "J" *dže*, and "W" *deblju*, which would be more normally rendered *ejč*, *džej*, and *dabəlju*. In Kubelka (1912b), the front matter presents an alphabet list that is different from that in the 1904 dictionary. While "H" and "J" are rendered with the same pronunciation as before, "W" is now *doblju*. The letter "A" is now given not as a diphthong but as *e*. Interestingly, in the digitized copy of Kubelka (1912b) that we were working from, some dictionary user of the past made a handwritten correction to the diphthong *ej* (see Figure 1).



ANGLEŠKA ABECEDA.

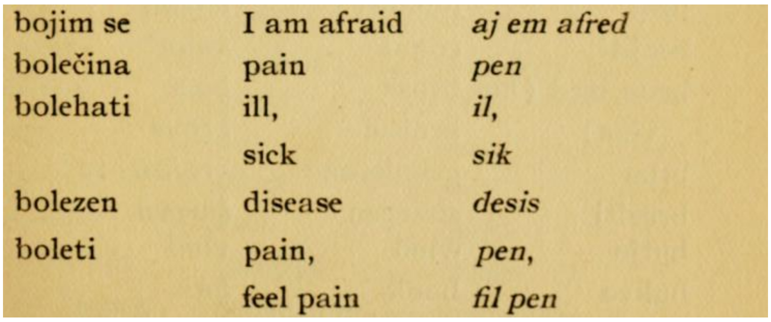
Črke	Izgovar-janje	Črke	Izgovar-janje	Črke	Izgovar-janje
a A	ej	j J	dže	s S	es
b B	bi	k K	ke	t T	ti
c C	si	l L	el	u U	ju
d D	di	m M	em	v V	vi
e E	i	n N	en	w W	doblju
f F	ef	o O	o	x X	eks
g G	dži	p P	pi	y Y	vaj
h H	eč	p Q	kju	z Z	zi.
i I	aj	r R	ar		

Figure 1: Kubelka 1912b (p. 11) with handwritten annotations

Kubelka (1904) presents a convoluted system to explain the pronunciation of American vowels in various contexts (front matter, pp. 12-14). The presentation format is a "rule", below which there is a chart in five columns containing a vowel, its pronunciation, an example of an English word with that vowel, the pronunciation of the sample English word, and finally the Slovenian translation of the English word. In Kubelka (1912b), while the greatly revised rule system for vowels is more accurate (and not rendered in charts), it is also quite long (front matter, pp. 12-22, 11 pages) and difficult to follow — and would have been entirely unsuitable for the general Slovenian immigrant audience it was intended for.

As an example, we can examine the discussion of the diphthong /eɪ/ in both of Kubelka's dictionaries. Kubelka (1904) discusses what he calls "compound vowels", which means words that have two contiguous vowel letters in their orthography. For the spelling "ai" as in *rain*, Kubelka lists the pronunciation *e* /ren/; for the spelling "ay" as in *pay*, he gives the pronunciation e^{i} /pej/ (p. 13). Certainly this is incorrect in terms of actual pronunciation; it is also confusing. Kubelka (1912b) has a different treatment that is closer to reflecting the actual pronunciation of /eɪ/ in different spellings. However, this treatment is dispersed over numerous disparate "rules" in this 11-page front matter section on the English vowels. For the spellings "ai" and "ay", Kubelka provides (among others) the examples *pay* /pej/ and *pain* /pejn/ (1912b, p. 18); for the spellings "ei" and "ey", he has (among other examples) *reign* /rejn/, *eight* /ejt/, and *grey* /grej/ (p. 19).

Within the body of the two dictionaries, we can see Kubelka's pronunciation rules in action. Recall that in both of his dictionaries, Kubelka took the unusual but useful step of providing pronunciation information about the English equivalents of his Slovenian lemmata. For the diphthong /eɪ/, Kubelka (1904) is consistent with the information given in his front matter; he uses for the lemma *bolečina* "pain" the pronunciation /pen/ (see Figure 2), for *dež* "rain" the pronunciation /ren/, for *plačati* "pay" the pronunciation /pej/, and for *siv* "gray" the pronunciation /grej/.



bojim se	I am afraid	*aj em afred*
bolečina	pain	*pen*
bolehati	ill,	*il,*
	sick	*sik*
bolezen	disease	*desis*
boleti	pain,	*pen,*
	feel pain	*fil pen*

Figure 2: Kubelka 1904 (p. 29)

On the other hand, the Slovenian–English component of Kubelka (1912b) is not consistent with what he said in the front matter about how /eɪ/ would be written; he renders this diphthong differently in closed syllables. For the lemma *dež* "rain", Kubelka gives the pronunciation /ren/, contradicting the front matter. The lemma *mučiti* has equivalents "to torture" /tu tortjur/ and "to pain" /tu pen/; the pronunciation of "pain" does not follow the front matter. The lemma *plačati* has "to pay" /tu pej/, in line with the front matter; *osem* "eight" /ejt/ and *siv-a-o* "grey" /grej/ (note the British English (BrE) spelling) are also in line with the front matter.

It is interesting to look at renderings of /ð/in the two Kubelka dictionaries. The preface of the 1904 work stated that /ð/ would be rendered as /dz/, and the 1912 front matter (consonant section) stated that /ð/ would be rendered /d/ or

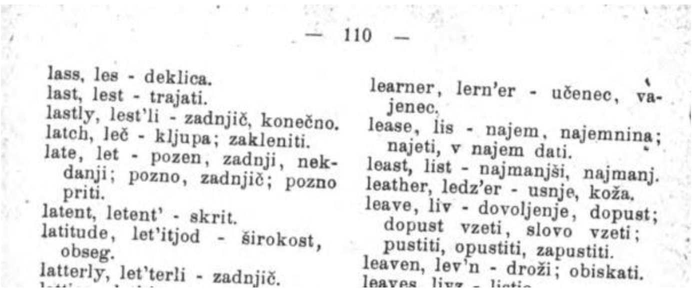
/t/. In 1904, for the lemma *oče*, the equivalent "father" is given with the pronunciation /fadzer/ (p. 74), consistent with Kubelka's (Slovenian-language) preface. However, for the lemma *usnje* "leather", Kubelka (1904) provides the pronunciation /leder/ (p. 110), in contradistinction to what his preface said. In Kubelka (1912b), the equivalent "father" is given the pronunciation /fadr/ (p. 354), consistent with what was stated in the front matter. The equivalent "leather" has the pronunciation /ledr/ (p. 411), also consistent. Our investigation showed that in Kubelka (1912b), the rendering of /ð/ is consistent throughout the dictionary and in line with what was stated in the front matter.

4.4 Košutnik 1912

Košutnik (1912) has a very short Slovenian-language preface (one and one-half pages). It is not signed with Košutnik's name but instead "The Publisher", although it is evident that this was written by the author. Košutnik discusses the pronunciation of /eɪ/, /æ/, /ɑ/, /ʌ/ /u/, and /yu/ as well as the pronunciation of /ɹ/ (the AmE alveolar approximant usually written as *r*). He notes that in the word *late*, the vowel is an *e* drawn out with a short *j* (IPA /y/) at the end. The word *latter* is pronounced as a kind of /ɑ/; however, this sound is not clear but is in between the IPA /ɑ/ and a short /e/. The word *lark* is pronounced with /ɑ/, as Košutnik calls it, "the true drawn-out *a*". About the /ɹ/ in *lark* he remarks that "this sound can only be heard before [sic] vowels and differs from Slovenian in that it is pronounced softly, without shaking the tongue". Note that in his discussion of *lark*, Košutnik is clearly referencing the AmE pronunciation /lɑɹk/ and not BrE /lɑːk/.

Košutnik's discussion of the pronunciation of *lug* /lʌg/ is rather obscure, but he does make the point that the /ʌ/ is "like a dark *a*", thus getting at the difficulty that this sound presents for non-native speakers of AmE. In contrast to the pronunciation of the vowel in *lug*, Košutnik notes that in the word *lune* there is a "true *u*"; elsewhere, in *lunula*, the pronunciation is *jo* (i.e., IPA /yo/). Taken as a whole, Košutnik's pronunciation observations from his preface are accurate. His perceptions of how AmE vowels differ from European ones are correct, even if his statements sometimes sound odd to modern ears.

Unlike the two Kubelka dictionaries, Košutnik (1912) has no front matter following the (Slovenian-language) preface, so the entirety of his discussion of pronunciation is in the one and one-half page preface devoted to vowels (as discussed above). In the English–Slovenian portion of the dictionary, we can find among Košutnik's English lemmata similar examples to those in Kubelka's two books. Recall that unlike Kubelka and more in line with standard bilingual dictionaries, Košutnik does not give pronunciation information in the Slovenian–English component of his dictionary, only in the English–Slovenian component. For /ð/ (not discussed in his preface), Košutnik's lemma *father* has the pronunciation /fa'dzer/. The lemma *leather* is pronounced /ledz'er/ (see Figure 3). These and other lemmata with /ð/ are rendered consistently; our investigation corroborated consistency throughout for /ð/. Note also that Košutnik gives word stress with ['], something that is entirely missing from Kubelka's two dictionaries.



— 110 —

lass, les - deklica.
last, lest - trajati.
lastly, lest'li - zadnjič, konečno.
latch, leč - kljupa; zakleniti.
late, let - pozen, zadnji, nekdanji; pozno, zadnjič; pozno priti.
latent, letent' - skrit.
latitude, let'itjod - širokost, obseg.
latterly, let'terli - zadnjič.
learner, lern'er - učenec, vajenec.
lease, lis - najem, najemnina; najeti, v najem dati.
least, list - najmanjši, najmanj.
leather, ledz'er - usnje, koža.
leave, liv - dovoljenje, dopust; dopust vzeti, slovo vzeti; pustiti, opustiti, zapustiti.
leaven, lev'n - droži; obiskati.

Figure 3: Košutnik 1912 (p. 110)

For the vowels discussed in the preface, it turns out that Košutnik is not entirely true to his word within the English–Slovenian portion of the dictionary. For example, for the diphthong /eɪ/, Košutnik says in the preface that it is an *e* drawn out with a short *j*, as in *late.* However, his pronunciation transcription uses *e* only:

eight /et/ (p. 87)
gray /gre/ (p. 99)
pain /pen/ (p. 119)
pay /pe/ (p. 120)
rain /ren/ (p. 127)
reign /ren/ (p. 128)

4.5 Kern 1919

While Kern 1919 has both an English and a Slovenian preface, each has different contents. In his English preface, Kern (1919) explains how he treats pronunciation: "In spelling and in pronunciation I have followed the latest authorities, particularly the *New Standard Dictionary,* which employs the phonetic method of indicating the pronunciation, and which is similar to the phonetic system of spelling and writing employed in the Slovene (Slovenian) language" (p. III). The *New Standard Dictionary* was first published in 1913 and revised through 1949; most likely Kern used the 1913 or 1914 edition. We did not have access to the *New Standard Dictionary*; however, its predecessor, *A Standard Dictionary* (Funk 1908) employed diacritics to indicate the different vowel sounds, most likely in a similar fashion to the *New Standard Dictionary* and most likely similar to Kern's system. For example, for the diphthong /eɪ/, the 1908 dictionary uses *e* with a macron over it, or *ē*. In the Slovenian preface, Kern adds that "Spelling and pro-

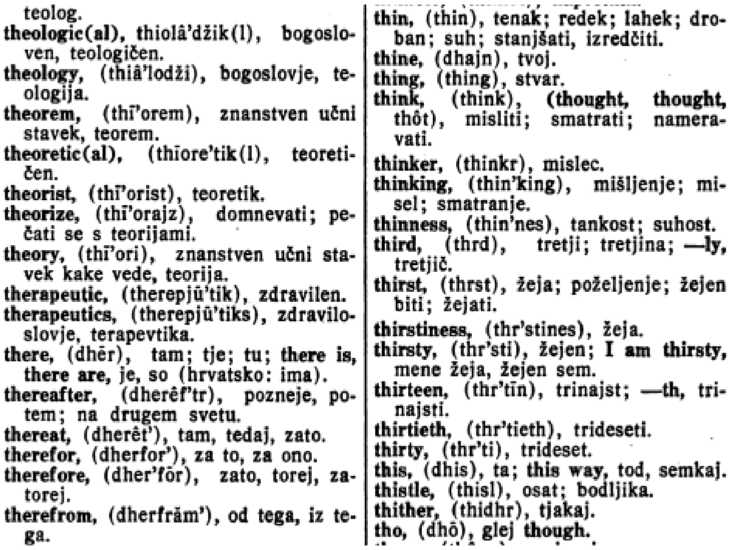
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1901 (Article)

nunciation of English words follows the American tradition and is somehow different from continental English".

Following Kern's two prefaces, we find a kind of user's guide [Pojasnila] in Slovenian (pp. VI-VII), which is mostly a pronunciation guide plus a brief note on compounds and a short explanation of how to use the dictionary (with an explanation of running heads and alphabetical arrangement of entries). On p. VI of this guide, the pronunciation advice covers vowels and three consonants: *th*, *dh*, and *w*. At the end (p. VII), Kern covers the pronunciation of other consonants.

Kern makes heavy use of diacritics to depict vowel sounds. As an example of one of his explanations, we can take the vowel /æ/ as in *cat*. Kern notes that ê is an open sound not found in Slovenian: "The mouth must be wide open horizontally and you should pronounce the letter e as in the word be-e-e-, as we say to imitate a sheep". As examples of English words, Kern provides *cat* /kêt/; *rat* /rêt/; *hat* /hêt/ (p. VI).

All the dictionaries examined so far mentioned /ð/ as in *the*, but none of them discussed /θ/ as in *thing*. Kern discusses and clearly differentiates the two sounds (see Figure 4). For /ð/, he uses *dh*, which he says has "a special sound" in English: "It is pronounced almost like *d*, if you press the tip of your tongue toward the teeth" (p. VI). As examples, Kern gives *that* /dhêt/, *this* /dhis/, and *father* /fādhr/. For /θ/, Kern uses *th*, "pronounced similarly to *dh*, only that instead of *d*, you try to pronounce *t*" (p. VI). English word examples provided are: *think* /think/, *thing* /thing/, and *nothing* /nă'thing/. Note that Kern's manner of indicating word stress (with a ['] after the stressed syllable) is the same as Košutnik's.



teolog.
theologic(al), thiolâ'džik(l), bogosloven, teologičen.
theology, (thiâ'lodži), bogoslovje, teologija.
theorem, (thī'orem), znanstven učni stavek, teorem.
theoretic(al), (thīore'tik(l), teoretičen.
theorist, (thī'orist), teoretik.
theorize, (thī'orajz), domnevati; pečati se s teorijami.
theory, (thī'ori), znanstven učni stavek kake vede, teorija.
therapeutic, (therepjū'tik), zdravilen.
therapeutics, (therepjū'tiks), zdraviloslovje, terapevtika.
there, (dhēr), tam; tje; tu; **there is, there are,** je, so (hrvatsko: ima).
thereafter, (dherêf'tr), pozneje, potem; na drugem svetu.
thereat, (dherêt'), tam, tedaj, zato.
therefor, (dherfor'), za to, za ono.
therefore, (dher'fōr), zato, torej, zatorej.
therefrom, (dherfrăm'), od tega, iz tega.
thin, (thin), tenak; redek; lahek; droban; suh; stanjšati, izredčiti.
thine, (dhajn), tvoj.
thing, (thing), stvar.
think, (think), **(thought, thought,** thôt), misliti; smatrati; nameravati.
thinker, (thinkr), mislec.
thinking, (thin'king), mišljenje; misel; smatranje.
thinness, (thin'nes), tankost; suhost.
third, (thrd), tretji; tretjina; **—ly,** tretjič.
thirst, (thrst), žeja; poželjenje; žejen biti; žejati.
thirstiness, (thr'stines), žeja.
thirsty, (thr'sti), žejen; **I am thirsty,** mene žeja, žejen sem.
thirteen, (thr'tīn), trinajst; **—th,** trinajsti.
thirtieth, (thr'tieth), trideseti.
thirty, (thr'ti), trideset.
this, (dhis), ta; **this way,** tod, semkaj.
thistle, (thisl), osat; bodljika.
thither, (thidhr), tjakaj.
tho, (dhō), glej **though.**

Figure 4: Kern 1919 (p. 241)

Once we turn to the lemmata in the English–Slovenian portion of the dictionary, we see that Kern is consistent in his notation of the sounds of English; the notations are in line with his discussion in the user's guide (Pojasnila):

> *eight* /ējt/ (p. 79)
> *gray, grey* /grēj/ (p. 107)
> *leather* /ledhr/ (p. 137)
> *pain* /pējn/ (p. 166)
> *pay* /pēj/ (p. 169)
> *rain* /rējn/ (p. 190)
> *reign* /rējn/ (p. 194)

5. Spelling: British versus American English

5.1 The two Kubelka dictionaries

Many divergences in BrE and AmE spelling took hold in the mid-1800s, and vocabulary differences between BrE and AmE also became more marked at this time (*Merriam-Webster* n.d.). Talkies, films with a soundtrack, emerged in the late 1920s and brought with them a greater international awareness of the distinctness of American accents. However, all of the dictionaries being examined here were produced earlier, during the silent film era, and like most of the world their authors appear not to have been very aware of, let alone versed in the differences between BrE and AmE, either in pronunciation or in spelling. Only Kern (1919) mentions the existence of two different Englishes when he notes that "Spelling and pronunciation of English words follows the American tradition and is somehow different from continental English" (Slovenian preface, p. IV). And Murphy (2018) homes in on the fact that the developing differences in the two Englishes were driven partially by the immigrants themselves:

> During the 19th century, the ethnic de-Britification of white America sped up, due to massive and increasingly diverse immigration. The children of immigrants to the US from Germany, Scandinavia, Ireland, and Italy became English speakers, but would never be Anglo-Saxons. (p. 64)

Google Ngram Viewer is a convenient, albeit imperfect tool (see Zhang 2015) for gaining insight into the development of spelling difference trends between BrE and AmE. In the 1830s and 40s, the use of the spellings *color*, *honor*, and *favor* began to become more frequent than the use of *colour, honour,* and *favour* in AmE (Google Ngram Viewer n.d.). Somewhat later, in the 1880s and 90s, *defence/defense* and *fibre/fiber* began to be distinguished in terms of frequency in AmE (Google Ngram Viewer n.d.). While *nitre/niter* appear to follow the pattern of *fibre/fiber*, because the former word from scientific terminology is so infrequent, the difference between the two variants is not as evident (Google Ngram Viewer n.d.). Despite these trends, Kubelka's 1904 Slovenian–English dictionary lists, under the lemma *barva,* the spelling "colour"; on the other hand,

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1901 (Article)

under the lemma *čast*, it lists "honor". In contradistinction to Kubelka (1904), in the Slovenian–English portion of the bidirectional Kubelka (1912b), the equivalent for the lemma *barva* is "color"; for *čast* it is "honor". While Kubelka (1904) does not contain a lemma *obramba*, Kubelka (1912b) does, and uses the BrE spelling in the equivalent "defence". This is logical if the divergence between *defence* and *defense* only began in the 1880s or 90s; the spelling *defense* would have been a newer phenomenon in 1912 than would have been the spellings *color, honor*, and *favor*. Likewise, in the English–Slovenian component of Kubelka (1912b), the dictionary settled on BrE spellings for the lemmas *defence, fibre*, and *nitre*. Note that the spelling choices in Kubelka (1912b) are consistent; the same spelling that appears in a headword in the English–Slovenian section will be used for an equivalent in the Slovenian–English portion of the dictionary.

5.2 Košutnik's dictionary

Unlike Kubelka and Kern, Silvester Košutnik was not an immigrant and compiled and published his dictionary in Ljubljana. Given the European setting, it would not be surprising for his choices to favor BrE spelling. Overall, it can be said that his preference is for the BrE spellings. For example, in the Slovenian–English portion of his bidirectional dictionary, Košutnik has "colour" (under the lemma *barva*), "honour" (under *čast*), "labourer" (under *delavec*), "to labour" (under *delati*). On the other hand, Košutnik (1912) has the equivalent "traveler" with an AmE spelling under the lemma *popotnik* (this is the only AmE spelling we found in the Slovenian–English section).

In the English–Slovenian part, Košutnik has the lemmata *colour, flavour, honour, labour,* and *labourer*, all with BrE spellings. In the English–Slovenian section, we found only the headword *favor* with an AmE variant; it is a mystery as to why Košutnik diverged from his normal pattern here. Additionally, as would be expected, Košutnik lists *fulfil* with a single *-l-*, in line with BrE spelling. Compare Košutnik's approach with that of Kern, below.

5.3 Kern's dictionary

As we noted, Kern (1919) is the only one of the three authors who explicitly mentions the differences between BrE and AmE. Unlike Košutnik, Kern overwhelmingly uses American spellings with rare forays into BrE (such as *defence)*; Kern is consistent in spelling the same word in the same way throughout the dictionary. Below, the sample list of lemmata provides insight into Kern's approach to spelling:

> *color*
> *councilor* (with equivalents "svetovalec, svetnik")
> *counselor* (with equivalent "svetovalec")
> *defence* with a cross-reference to *defense* [no indication that *defence* is BrE]

disfavor
dishonorable
encyclopedic(al)
endeavor
favor
favorable
favorite
fervor
flavor
neighbor
neighborhood
neighborly
tumor

Kern does provide both British and American versions of a lemma in some infrequent instances, rather than using a cross-reference (as he did for *defence*, given above). Under the letter "E", Kern gives the lemma *edema, oedema*, with the AmE variant first and the BrE spelling second, although this choice could be due to considerations of the alphabetical order. Under the letter "O", there is no listing for *oedema* (so obviously no cross-reference to *edema*). Interestingly, if we can rely on Google Ngram Viewer (n.d), *edema* and *oedema* were used about equally in AmE in 1900, around the time Kern's dictionary was made. While Kern's treatment of *edema, oedema* could be motivated, the same cannot be said for his handling of *eon, aeon*. The lexicographic treatment of *eon, aeon* is exactly the same as that of *edema, oedema*: under the letter "E" we have both spellings *eon, aeon*; under the letter "A" there is no listing (and hence no cross-reference). Google Ngram (n.d.) seems to indicate that *aeon* (in contrast with *eon*) was not a viable variant; at that time, it appears to have been used rarely in both BrE and AmE. It is possible that Kern, a medical doctor, knew the *aeon* spelling through his knowledge of Latin; his education both at a gymnasium in Ljubljana and at a seminary in Minnesota would have exposed him to Latin frequently. It is also possible that he was influenced by one or more monolingual dictionaries of that time. For example, *Webster's New International Dictionary*, 1st edition (Harris 1909), has the headword *æon* as well as *eon*. In both places, the variant spelling is given and a full treatment of the meaning appears.[5] Note that Kern's own treatment in his bilingual dictionary is more modern than that in the monolingual Webster's that gave equal treatment to *aeon* and *eon*.

As for *esthetic*, Kern lists it as a headword with an equivalent and with a cross-reference to *aesthetic*. At the lemma for *aesthetic*, the same equivalent is listed that was available at *esthetic*. However, at *aesthetic*, there is no cross-reference back to *esthetic*. Neither word has any label for BrE or AmE. This treatment could indicate that Kern considered *aesthetic* as the main variant (so that it did not need a cross-reference back to *esthetic*); apparently he considered *esthetic* to be important enough to list an equivalent in addition to the cross-reference to *aesthetic*. If this was indeed Kern's reasoning (which we cannot know), then Google Ngram

Viewer (n.d.) would seem to support that. The Ngram Viewer indicates that in both North America and Britain, the variant *aesthetic* has always been the more frequent spelling. Today, dictionaries in both North America and Britain list *aesthetic* as the main variant, though *esthetic* is included as the second variant or marked as North American.

While Kern was consistent in spelling the same word in the same way throughout the dictionary, he was not always consistent in his treatment of related phenomena of AmE/BrE spelling differences. It is not clear that he should have been consistent, since different lexical items, even with similar forms, can have different timelines for when a spelling variant begins to predominate. For example, many verbs and derivative nouns have *-l-* in BrE and *-ll-* in AmE; examples are *enrol/enroll, enrolment/enrollment, fulfil/fulfill, fulfilment/fulfillment,* and *instal/install, instalment/installment*. Below are the choices Kern made for the headwords in his English–Slovenian dictionary:

enroll
enrollment
fulfil
fulfilment
install
instal(l)ment

While some of Kern's choices follow AmE spelling (*enroll*, *enrollment*; *install*), some follow BrE (*fulfil, fulfillment*). The lemma *instal(l)ment* has both British and American spellings, but without comment. When we examine the frequency graphs of Google Ngram Viewer (n.d.), we discover that in the time frame in which Kern's dictionary was being created (1905–1919; cf. Javh-Kern 1937), each of the words above had somewhat different usage arcs. For example, the pair *fulfill* and *fulfil*, as well as the pair *fulfillment* and *fulfilment* were apparently used approximately equally in the American context from the early 1870s to around 1915, at which time *fulfill* and *fulfillment* began slowly to be used more frequently. *Instalment* and *installment* appear to have been on more or less equal footing until the mid1840s, with an increase in the use of *installment* from then on. This tells us that all of these variants (except *instal*) were circulating in usage during Kern's time; it would have been impossible for him to determine which ones were most frequent. Given this, it is not clear why Kern chose to put one *l* of *instal(l)ment* in parentheses.

6. Discussion and conclusions

6.1 Background and experience

Despite sustained investigation, many aspects of the background of our three dictionary authors remain obscure. This is unfortunate, because such information could shed light on what skill sets they possessed and how their knowledge and experiences motivated them to create their dictionaries and prepared them to do this successfully.

From what we have been able to determine, Fran/Frank Javh-Kern was the most educated of the three authors, with a gymnasium education in Ljubljana, and a seminary education in Minnesota followed by a degree in medicine. Kern was a practicing physician and a prominent member of the Cleveland, Ohio community (Javh-Kern 1937). In his later years, he taught college courses in Cleveland.[6] In addition to Slovenian and English, Kern certainly knew German and Latin.

Silvester Košutnik was the son of a teacher and a teacher himself, but like most teachers at that time probably did not receive education after secondary school. Košutnik was also a prolific writer and translator into Slovenian. Notably, he translated *Uncle Tom's Cabin* into Slovenian — from a German version rather than the English original (Hladnik 1985; Trupej 2015). His translation work would have given him familiarity with both monolingual and bilingual dictionaries. Apart from Slovenian, he was familiar with English, French, German, Italian, and Serbo-Croatian (Hladnik 1983).

Viktor/Victor Kubelka's background is the most mysterious of the three.[7] He was born in Ljubljana, but no information about his formal education is available; he apparently did not study in the U.S. It appears that he knew German, Czech, and Croatian in addition to English and Slovenian. He worked as an inspector in the War Department for U.S. Military Intelligence during World War I. For a few years after the war he worked as a director of the Commercial Department of the Czechoslovak Consulate General; his father's native language was Czech. In sum, all three authors had vast experience with languages, Kern had higher education and Košutnik and Kern pursued intellectual endeavors. All of the authors had the zeal to provide the necessary dictionary tools to help their compatriots — Slovenian immigrants or those getting ready to emigrate — master English.

As far as dictionary-writing is concerned, all three authors were outright amateurs as well as groundbreaking pioneers. None of them were linguists; despite their collective knowledge of many languages, they had no experience working with language analytically. During the time they were working on their dictionaries, there were simply no resources on English grammar, pronunciation, etc. written in Slovenian.[8] So, they had to invent the wheel. First, they had to grasp what the proper explanations were for various language phenomena; once they had come up with an explanation, they had to discern how best to present it for their intended audience. For example, as non-linguists they had to describe the physical movements of the tongue during the pronunciation of different phonemes. Above, we gave a few examples of Kern's fairly successful descriptions. While he may have been less successful, Kubelka was certainly not lacking in energy, when he advises (in his 1912b Slovenian front matter, p. 26) how to deal with /ð/:

> … is pronounced if you push the tip of your tongue toward the upper teeth and try to pronounce *ds*. They are marked in this book by *d* and *t*.

For the various English phonemes, the authors had to ascertain which Slovenian words had the same or similar phonemes, so that they could draw parallels between the two languages. It is difficult today to fathom how they achieved what they did. The authors also had to learn and sometimes invent Slovenian linguistic terminology. While their efforts with grammar and grammatical terms will be addressed in a future article, even in the realm of pronunciation, they had to grapple with what kind of consonants are /θ/ and /ð/, and what is a /w/ or an /ɹ/ (the AmE alveolar approximant written as *r*) and how exactly these sounds work in the English language. These are concepts that linguists learn during their formal education and then, depending on their area of specialization, could end up studying over many years.

6.2 The state of the art

In terms of its level of professionalism, bilingual lexicography has always been at least a step or two behind monolingual lexicography, despite the fact that bilingual dictionaries existed first. Héja, Lipp and Prószéky (2023) maintain that, while the first bilingual wordlist appeared around 2400 BCE, only in the 1950s — well after Kubelka's, Košutnik's and Kern's time — did real scientific discussion of bilingual lexicography begin. To the challenges faced by the monolingual lexicographer in how to discriminate meaning and present linguistic information cogently, the bilingual lexicographer must add the demand of juxtaposing the semantic units of two languages that never or rarely match in their equivalence. This requires a degree of thorough knowledge of both languages that usually does not reside within a single individual working on a bilingual dictionary. Bilingual dictionaries are always born in situations of language and cultural contact (cf. Béjoint 2016 and Fontenelle 2016) where the dire need outweighs the qualms of imperfection. If the lack of ideal qualifications for the job had prevented our three authors from creating their books for the benefit of the Slovenian immigrant population, as Voltaire said, the best would indeed have been the enemy of the good.

6.3 Understanding the target audience

Who was the target audience of our three dictionary makers? To answer this question, we must first understand the education and schooling that Slovenian immigrants were likely to have had in the home country. From December 1774 in the Hapsburg Monarchy (of which Slovenian territories were a part), primary school was obligatory for children aged six to 12.[9] While it was not dictated by law, it was presupposed that the language of instruction would be German, with the understanding that pupils in Slovenian territories would need supplementary mother tongue support (Okoliš 2008: 45). From 1809 to 1813, the Slovenian language was introduced in schools in some of the Slovenian territories (Okoliš 2008: 51). As the 19th century progressed, the use of Slovenian as the language of instruction gained momentum in all of the Slovenian territories (Fedor 2017). In 1869, compulsory primary schooling was standardized and increased from six to eight years (Federal Ministry of Education, Science and Research, Republic of Austria, n.d.). While the territories that were inhabited by Slovenian speakers were more impoverished than other areas of the Hapsburg Monarchy, still we can surmise that most immigrants would have had a minimum of four years of primary education and would have been able to read, even if many of them did not use their literacy skills often. In this respect, they were similar to native-born Americans who likewise had high levels of literacy at this time (Lynch 2011).

All three of our dictionary authors knew their audience. Košutnik never emigrated and had to have been (as the son of a teacher and a teacher himself) very familiar with the educational level of potential emigrants to the New World; we have already seen that he references "those layers of society for which this booklet is intended" (1912, p. 4). Kubelka immigrated to the U.S. as a young adult; while the exact year he arrived has yet to be determined, he was apparently in his early twenties — and thus had experienced the education system in Ljubljana firsthand (although we do not know whether he had any secondary education). Prior to immigration, Kern received both primary and secondary (gymnasium) education in Slovenian territories and then went on to receive higher education in the U.S. From the age of 16 he lived in the U.S. among Slovenian immigrants. He states in the introduction to his memoir: "I had an unusual opportunity to observe the course and development of Slovenian colonization in America …", and "[i]n my public activity the good of the Slovenian people here and in the old country was my main goal, only then my own benefit" (Javh-Kern 1937). While Kern was far from being a typical Slovenian immigrant of that time, there is no question that he knew his compatriots well.

Despite having knowledge of their audience, as amateur dictionary writers, our authors were not fully successful in designing books that would convey pronunciation and spelling information effectively. While their books were used — and treasured in the families of Slovenian immigrants — it is clear that their guidance would often have been obscure or completely opaque. As the examples above from the prefaces and front matter show, the authors did their best to explain pronunciation and demonstrated their own insight into pronunciation issues. However, most of the information they provided would not have been unusable even for astute and motivated learners. Of the three authors, Kubelka has the most information about pronunciation in his prefaces and front matter, and it is the least usable. Only Kern comments on the difference between AmE and BrE spelling in his (Slovenian-language) preface, but it is unlikely that this was a significant topic for his users.

It is not a great tragedy that the three authors toiled over prefaces and front matter that was, most likely, completely ignored. This has been the fate of most lexicographers everywhere, in all times. The more important question is whether the pronunciation and spelling guidance within the body of the actual dictionary was successful. To begin with pronunciation, it cannot be overemphasized how important it is for the non-native speaker. While native speakers do not think much about pronunciation, the non-native learner can be obsessed with it since a mispronunciation can render a person incomprehensible and cause complete communication failures. What is more, due to the historical nature of the English spelling system, pronunciation is simply not available to a language learner from a written text alone. Spelling is less important, as the Slovenian immigrants in the mines and the steel mills did not need to write much. What is more, in this era most average Americans, native and non-native speakers alike, were not aware that there were BrE and AmE variants.

How did the three authors do with the pronunciation guidance in their dictionaries? Our review must be mixed. All of them used the only thing available to them at that time, systems of respelling in a more phonetic manner. Apparently, Kubelka and Košutnik invented their systems. The systems they created, due to these two authors' difficulty with some of the challenging sounds of English (/ɹ/, /ð/, /θ/, /eɪ/, /æ/, /ɑ/, /ʌ/, /eɪ/, etc.), could not provide fully effective guidance to users. In addition, as noted above, Kubelka and Košutnik did not use their respelling systems consistently. On the other hand, Kubelka did his readers a service by his (unorthodox) use of pronunciation guides next to the equivalents in his two Slovenian–English dictionaries. This technique did not catch on in most bilingual dictionaries intended for encoding, but perhaps it should have.

Kern differs from Kubelka and Košutnik in that his respelling system for pronunciation was either taken whole cloth or adapted from the respelling system in one of the early editions of the *New Standard Dictionary*. This was a good choice, given that Funk and Wagnalls dictionaries aimed to adhere to accurate phonetics. For this reason, Kern's system is the most consistent, so it might have been more useful to the Slovenian immigrants exposed to it.

Above we indicated that in terms of spelling choices, Kubelka (1904) was inconsistent, Kubelka (1912b) was consistent in representing AmE, and Košutnik was consistent in representing BrE. We have noted that, of the four books examined, Kern's dictionary was the most professionally done. In terms of spelling, Kern was consistent overall and primarily used AmE variants. As we speculate above, when Kern does use BrE spellings, there is some (albeit limited) evidence that he does so in cases where the BrE variant was still widely circulating in the society along with an AmE variant.

6.4 Final musings

The dictionaries of Kubelka, Košutnik and Kern (as well as other reference books of the Slovenian immigration to the U.S.), despite their flaws and inconsistencies, were of immense value to the population for whom they were intended. These books have not been forgotten. The descendants of Slovenian–American immigrants know these dictionaries or still have them in their families; they reside in libraries and museums in the U.S. and in Slovenia. They also live on in this ongoing project to document their lexicographic value and to augment the lexi-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1901 (Article)

cographic knowledge base of our field. More work is needed for a full understanding of what the new bilingual Americans of the early 20th century thought about language and dictionaries.

Acknowledgements

We are grateful to the Slovenian Research Agency, for providing funding for the grants *Slovarji in izkušnje Slovencev–emigrantov v Združenih državah Amerike (konec 19. in začetek 20. stoletja)/Dictionaries and the Slovenian Immigrant Experience in the United States (Late 19th, early 20th Century)* [BI-US/22-24-042] and *Stare besede, nove besede, novi svet: Življenje slovenskih leksikografov — priseljencev v ZDA/Old Words, New Words, New World: The Lives of Slovenian Immigrant Lexicographers in the U.S.* [BI-US/22-24-030].

We also would like to thank those who assisted us during this project: Georgene Agnich, Lauren Baltz, Cynthia Mencin Barkowski, Jeffrey Dormish, Geraldine Hopkins, Rose Marie Macek Jisa, Lorraine M. Kaup, Laura Kortz, Tom Percic, April Sleyko, Deacon John P. Vidmar, Ph.D., and Mary Lou Deyak Voelk. To these very generous people, we say *Vsem iz srca hvala!* May your efforts to foster Slovenian culture in the U.S. continue.

Endnotes

1. This was reprinted in 1903 and 1908 with the title *Angleščina brez učitelja v slovenskem jeziku. Pomočna knjiga za potovalce v Ameriko* [English without a Teacher in the Slovenian Language. A Handbook for Travelers to America]. The 1908 edition is available on Google Books.
2. See Kalc et al. (2020) concerning the years of Slovenian immigration to the U.S., 1890–1914.
3. The earliest German–English lexicographic source is a six-language bilingual dictionary published in the first half of the 16th century, and there are other German–English sources as well (McLelland 2018). By the 19th century, as was the case with other language pairs in Europe, the number of English–German bilingual dictionaries had increased. However, it is doubtful that our authors were consulting these or other dictionaries in early 20th-century America.
4. *Slovensko–angleški razgovori, Slovenian–English Interpreter* (1912a); this work is not analyzed here.
5. Later editions of *Webster's New International Dictionary* (2nd edition, Neilson 1934 and 3rd edition, Gove 1961) continued to have full entries under *aeon* as well as under *eon*.
6. Personal communication, Deacon John P. Vidmar, Ph.D., Slovenian Catholic Mission, Lemont, Illinois.
7. Information about Viktor Kubelka was gleaned from research on Ancestry.com and at the Slovenian Genealogy Society, International, Inc. in Cleveland, Ohio.
8. The first readily available grammar (with pronunciation explanation) would have been the immigrant Jeram's English grammar in Slovenian (1895). We can be certain that Kern was familiar with this book (see his memoir, Javh-Kern 1937), but we do not know whether the other authors knew or used it.
9. The government of Austria (under Empress Maria Theresa) instituted a requirement for six years of primary education (Cvrček 2020).

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1901 (Article)

References

Note: Citations of Kern are listed as they appear in the respective publication, as Kern or as Javh-Kern.

A. Dictionaries

Funk, Isaac K. (Ed.). 1908. *A Standard Dictionary of the English Language.* New York/London: Funk & Wagnalls Company.

Funk, Isaac K. (Ed.). 1913. *New Standard Dictionary of the English Language*. New York/London: Funk & Wagnalls Company.

Gove, Philip Babcock (Ed.). 1961. *Webster's Third New International Dictionary.* Third edition. Springfield, Mass.: G. & C. Merriam Company.

Harris, William Torrey (Ed.). 1909. *Webster's New International Dictionary.* First edition. Springfield, Mass.: Merriam–Webster Inc.

Kern, Frank J. 1919. *A Complete Pronouncing Dictionary of the English and Slovene Languages for General Use. Popoln angleško-slovenski besednjak z angleško izgovarjavo.* Cleveland, Ohio: Tisk ameriške domovine. (Kern 1919)

Košutnik, Silvester. 1912. *Ročni slovensko–angleški in angleško slovenski slovar: Zlasti namenjen izseljencem v Ameriko.* Ljubljana: Anton Turk. (Košutnik 1912)

Kubelka, Victor J. 1904. *Slovensko–angleški žepni rečnik v olajšavo naučenja obeh jezikov. Slovenian–English Pocket Dictionary to Facilitate the Study of Both Languages.* New York: Victor J. Kubelka. (Kubelka 1904)
https://babel.hathitrust.org/cgi/pt?id=loc.ark:/13960/t4qj8p026&view=1up&seq

Kubelka, Viktor J. 1912a. *Slovensko–angleški razgovori. Slovenian–English Interpreter*. New York: Press of Stettiner Bros. (Kubelka 1912a)
https://hdl.handle.net/2027/nyp.33433021125889

Kubelka, Viktor J. 1912b. *Slovensko–angleška Slovnica, Tolmač, Spisovnik in Navodilo za Naturalizacijo. Angleško-Slovenski in Slovensko–Angleški Slovar. Slovenian–English Grammar Interpreter, Letterwriter and Information on Naturalization. English–Slovenian and Slovenian–English Dictionary.* First edition. New York: Viktor J. Kubelka. (Kubelka 1912b)

Merriam-Webster. n.d. Is it 'Autumn' or 'Fall'? Why Does this Season Have Two Vastly Different Names?
https://www.merriam-webster.com/words-at-play/autumn-vs-fall [6 July 2023].

Neilson, William Allan (Ed.). 1934. *Webster's New International Dictionary*. Second edition. Springfield, Mass.: G. & C. Merriam Company.

B. Other literature

Adams, Michael. 2015. Language Ideologies and *The American Heritage Dictionary of the English Language*: Evidence from Motive, Structure, and Design. *Dictionaries: Journal of the Dictionary Society of North America* 36: 17-46.

Anonymous. 1895. *Angleščina brez učitelja. Pomočna knjiga za izseljence.* Ljubljana: Josip Pavlin.

Anonymous. 1903, 1908. *Angleščina brez učitelja v slovenskem jeziku. Pomočna knjiga za potovalce v Ameriko.* Ljubljana: Josip Pavlin. https://books.google.com/books?id=ksWCvwEACAAJ&printsec=frontcover&hl=sl#v=onepage&q&f=false

Bajt, Drago (Ed.). 1985. *France Prešeren v prevodih.* Zbornik Društva slovenskih književnih prevajalcev 8/9. Ljubljana: Društvo slovenskih književnih prevajalcev.

Béjoint, Henri. 2010. *The Lexicography of English*. Oxford: Oxford University Press.

Béjoint, Henri. 2016. Dictionaries for General Users: History and Development; Current Issues. Durkin, Philip (Ed.). 2016: 7-24.

Cvrček, Tomáš. 2020. *Schooling under Control: The Origins of Public Education in Imperial Austria 1769–1869*. Tübingen: Mohr Siebeck.

Durkin, Philip (Ed.). 2016. *The Oxford Handbook of Lexicography*. Oxford: Oxford University Press.

Engelberg, Stefan, Heidrun Kämper and Petra Storjohann. (Eds.). 2018. *Wortschatz: Theorie, Empirie, Dokumentation*. Berlin/Boston: De Gruyter.

Federal Ministry of Education, Science and Research, Republic of Austria. n.d. *History of the Austrian School System.* https://www.bmbwf.gv.at/en/Topics/school/school_syst/hist_school_syst.html [12 July 2023].

Fedor, Helen. 2017. 19th-Century Slovenian Primers and Readers. *Library of Congress Blogs,* 30 January 2017. https://blogs.loc.gov/international-collections/2017/01/19th-century-slovenian-primers-and-readers/#:~:text=The%20use%20of%20Slovenian%20in,creation%20of%20Sunday%2Donly%20schools [12 July 2023].

Fontenelle, Thierry. 2016. Bilingual Dictionaries: History and Development; Current Issues. Durkin, Philip (Ed.). 2016: 44-61.

Google Ngram Viewer. *aesthetic, esthetic; color, colour; defense, defence; edema, oedema; eon (British English); eon (American English); favor, favour; fiber, fibre; fulfil, fulfill, fulfilment, fulfillment; honor, honour; instalment, installment; niter, nitre* [July 2023].

Héja, Enikő, Veronika Lipp and Gábor Prószéky. 2023. Bilingual Lexicography. *Oxford Bibliographies,* 21 February 2023. DOI: 10.1093/OBO/9780199772810-0301 [12 July 2023].

Hladnik, Miran. 1983. Tipi slovenske trivialne proze na začetku tega stoletja. Zadravec, Franc, Franc Jakopin and France Bernik (Eds.). 1983: 125-136.

Hladnik, Miran. 1985. Svobodno po nemškem poslovenjeno: Popularni prevedeni žanri 19. stoletja. Bajt, Drago (Ed.). 1985: 191-199.

Jackson, Howard. 2002. *Lexicography: An Introduction*. London/New York: Routledge.

Javh-Kern, Frank. 1937. *Memoirs on the Occasion of the Thirtieth Anniversary of Arrival in America*. [Typescript read at the Slovenian Genealogy Society International, Inc.]

Jeram, Peter Jos. 1895. *Slovensko–angleška slovnica*. Tower, Minn.: Amerikanski Slovenec.

Kalc, Aleksej, Mirjam Milharčič Hladnik and Janja Žitnik Serafin. 2020. *Doba velikih migracij na Slovenskem*. Ljubljana: Založba ZRC.

Landau, Sidney I. 2001. *Dictionaries: The Art and Craft of Lexicography*. Second edition. Cambridge: Cambridge University Press.

The Library of Congress. n.d. *Immigrants in the Progressive Era.* (U.S. History Primary Source Timeline).
https://www.loc.gov/classroom-materials/united-states-history-primary-source-timeline/progressive-era-to-new-era-1900-1929/immigrants-in-progressive-era/#:~:text=Between%201900%20and%201915%2C%20more,the%20previous%2040%20years%20combined [14 July 2023].

Lynch, Jack. 2011. "Every Man Able to Read": Literacy in Early America. *CW Journal,* Winter 2011.
https://research.colonialwilliamsburg.org/Foundation/journal/Winter11/literacy.cfm [12 July 2023].

McLelland, Nicola. 2018. Deutsch als Fremdsprache und die deutsch–englische Lexikographie bis 1900. Engelberg, Stefan, Heidrun Kämper and Petra Storjohann (Eds.). 2018: 295-320.
https://doi.org/10.1515/9783110538588.

Mulaček, Ivan. 1930. *Učbenik angleškega jezika*. Ljubljana: Jugoslovanska knjigarna.

Murphy, Lynne. 2018. *The Prodigal Tongue: The Love–Hate Relationship between American and British English.* New York: Penguin Books.

Okoliš, Stane. 2008. *Zgodovina šolstva na Slovenskem.* Ljubljana: Slovenski šolski muzej.
https://stanko-okolis.si/wp-content/uploads/2018/05/Zgodovina-%C5%A1olstva-na-Slovenskem.pdf

Passy, Paul. 1904. [Supplement] Aim and Principles of the International Phonetic Association. *Le Maître Phonétique* 19: 1-20.
http://www.jstor.org/stable/44703664 [8 July 2023].

The Principles of the International Phonetic Association. 1912. [Supplement] *Le Maître Phonétique* 27: 1-40.
http://www.jstor.org/stable/44707964 [8 July 2023].

Shapiro, Rebecca. 2020. Late Eighteenth-century English Orthoepic Dictionary Front Matter. *Lexicography Asialex* 7(1–2): 103-114.
https://link.springer.com/journal/40607/volumes-and-issues/7-1

Stanonik, Janez. 1996. Slovenci v Združenih državah: obdobje 1848–1891. *Dve domovini* 7: 113-129.

Trupej, Janko. 2015. Recepcija štirih ameriških romanov in njihovih slovenskih prevodov v luči ideologije rasizma. *Primerjalna književnost* 38(2): 213-235.

United States Census Bureau. n.d. *1860 Fast Facts.*
https://www.census.gov/history/www/through_the_decades/fast_facts/1860_fast_facts.html [8 July 2023].

Zadravec, Franc, Franc Jakopin and France Bernik (Eds.). 1983. *Obdobje simbolizma (Obdobja 4).* Ljubljana: Filozofska fakulteta.

Zakrajšek, Kazimir. 1923. *Abecednik za ameriške Slovence*. Second edition. Chicago: Edinost.

Zhang, Sarah. 2015. The Pitfalls of Using Google Ngram to Study Language: Garbage in, Garbage out When It Comes to Big Data Analysis of Language and Culture. *Wired,* 12 October 2015.
https://www.wired.com/2015/10/pitfalls-of-studying-language-with-google-ngram/ [10 July 2023].

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1902 (Article)

Algorithmic Complexity and Learnability in German Monolingual Learner Lexicography. A Case Study

Alberto Galván-Santana, *Department of Applied Linguistics,*
Universitat Politècnica de València (UPV), Spain
(algalsan@doctor.upv.es) (https://orcid.org/0000-0002-4665-5956)

Abstract: This paper analyzes the algorithmic complexity (also known as Kolmogorov complexity or descriptive complexity) of the lemma corpus included in the *Wortfamilienwörterbuch der deutschen Gegenwartssprache* (WfWG; Augst 2009) as a function of its macrostructural arrangement. The results show that, compared to the alphabetical order, the WfWG word-family arrangement produces an algorithmically more compressible, and therefore less complex version of the lemma corpus. This observation points to a higher degree of learnability and cognitive accessibility of the lemma corpus arranged in word families.

Keywords: MONOLINGUAL LEARNER'S DICTIONARY, MACROSTRUCTURE, NAVIGATION, LEARNABILITY, ALGORITHMIC COMPLEXITY, COMPRESSION

Zusammenfassung: Algorithmische Komplexität und Lernbarkeit in der einsprachigen Lernerlexikographie des Deutschen. Eine Fallstudie. Dieser Beitrag analysiert die algorithmische Komplexität (auch Kolmogorow-Komplexität oder Beschreibungskomplexität) des im *Wortfamilienwörterbuch der deutschen Gegenwartssprache* (WfWG; Augst 2009) enthaltenen Lemmakorpus in Abhängigkeit von dessen makrostruktureller Anordnung. Die Ergebnisse zeigen, dass die Wortfamilienanordnung im Vergleich zur alphabetischen Reihenfolge eine stärker komprimierbare und daher weniger komplexe, d. h. kürzer beschreibare Version des WfWG-Lemmakorpus darstellt. Diese Beobachtung deutet auf einen höheren Grad an Lernbarkeit und kognitiver Zugänglichkeit des in Wortfamilien angeordneten Lemmakorpus hin.

Schlüsselwörter: EINSPRACHIGES LERNERWÖRTERBUCH, MAKROSTRUKTUR, NAVIGATION, LERNBARKEIT, ALGORITHMISCHE KOMPLEXITÄT, KOMPRESSION

1. Introduction

The increasing use of the monolingual learner's dictionary (MLD) in the field of L2 acquisition has created a growing demand for a psycho-cognitive motivation in the conception of lexicographic texts that would facilitate an intrinsic activation of grammatical knowledge in the dictionary user (Fuertes-Olivera and

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1902 (Article)

Tarp 2011; Haß-Zumkehr 2012; Kövecses and Csábi 2014; Kremer et al. 2008).

Nowadays the psycho-cognitive approach in the elaboration of MLD is based on different models and cognitive theories focused on semantic aspects: frame semantics (Fillmore 1982, 1985), the theory of conceptual metaphor (Lakoff 1993; Lakoff and Johnson 1980), or the model of principled polysemy (Evans 2009; Tyler and Evans 2004), among others. With regard to the microstructure, recent approaches have led to the formulation of a "cognitive lexicography" (Ostermann 2015), focused on the applicability of cognitive linguistics to the propositional format of the lemma definition. Regarding the macrostructure, the application in the area of e-lexicography of these linguistic-cognitive theoretical frameworks has resulted in the creation of lexical knowledge databases such as WordNet (Miller et al. 1990), MindNet (Richardson et al. 1998), FrameNet (Fillmore et al. 2003), and HowNet (Dong and Dong 2006).[1] In print lexicography, the application of various linguistic knowledge and theories to lexicography has produced a variety of macrostructural designs whose intended goals can be reduced to a single common denominator: assisting the L2 student in the acquisition of language production and comprehension skills. This (primarily) printed dictionary type is represented by onomasiological dictionaries (Casares 2013; Dornseiff 2020; Rey-Debove and Rey 2009; Simone 2010), collocation and combinatorial dictionaries (Benson et al. 2009; Bosque 2006; Häcki Buhofer et al. 2014; Mel'čuk et al. 1999), and word-family dictionaries (Augst 2009; Davau et al. 1984; Kirkpatrick 1983), among others.[2]

In this context, the general objective of this project is to promote the development of macrostructural arrangement criteria capable of facilitating language acquisition in MLD users through their involvement in cognitive-efficient inference processes. This project is thus conceived as a contribution aiming to close the aforestated research gap — the demand for psycho-cognitive approaches in monolingual learner lexicography — with the lexicographic macrostructure at the center of the analysis.

To this end, we propose to address the psycho-cognitive relevance of the macrostructure from an extralinguistic perspective: the Algorithmic Information Theory (AIT; Chaitin 2004; Grünwald and Vitányi 2008). This information-based approach will allow us to analyze the cognitive accessibility of the lemma corpus in terms of algorithmic complexity (AC; Kolmogorov 1963; Chaitin 1969; Solomonoff 1964a, 1964b). The notion of AC is inversely correlated with learnability so that the learnability of any given data set increases with decreasing complexity of its structural organization (Clark and Lappin 2013; Fulop and Chater 2013; Kempe et al. 2015; Zenil and Gauvrit 2017).[3] In this regard, we argue that language learning with the help of a (monolingual learner's) dictionary, in this case the *Wortfamilienwörterbuch der deutschen Gegenwartssprache* (WfWG), can generally be conceived of in computational terms as a supervised learning task, in which lower complexity of the data structure (the lemma corpus) expedites the path for the learning algorithm (the dictionary user) to efficiently approximate the program (the grammar) that generates the data (cf. Grünwald 2005: 7-10).

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1902 (Article)

In accordance with this methodological approach, the overall objective of this study translates into the following specific objectives:

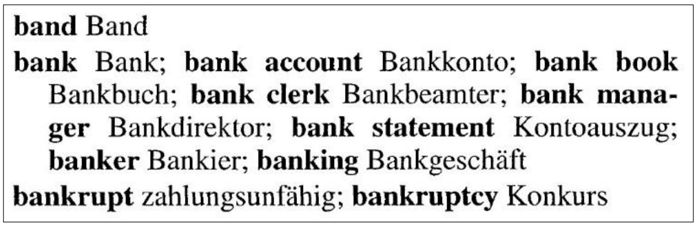
(O1) determine whether the macrostructural arrangement of the WfWG lemma corpus has an impact on its AC value;
(O2) evaluate to what extent the AC value of the WfWG corpus varies as a function of its macrostructural order, in word families and alphabetical, respectively;
(O3) identify, among the aforementioned ordering criteria, the macrostructural arrangement leading to a lower AC value to the corpus.

1.1 The lexicographic macrostructure

In general terms, the word macrostructure refers to the external structure (of a vertical or paradigmatic nature) that relates to the lemma corpus and the ordered representation of its elements (Engelberg and Lemnitzer 2009). The most elementary version of an alphabetic macrostructure is the simple alphabetical order (Gouws 2003). In a simple alphabetic dictionary, the number of indexed items, which are always main lemmata, is equal to the number of dictionary articles (Wiegand 1989). This implies that all lemmata are ordered according to the alphabetic value of the first character of the lemma sign and, consecutively, according to the alphabetic value of the following characters. Martínez de Sousa (2009: 214) defines this lexicographic procedure as "simple or lexicological" alphabetical arrangement.

In this regard, an alternate alphabetical arrangement procedure is characterized by the presence of groupings of sublemmata (Gouws 2003). These groups — integrated and hierarchically subordinated to the main lemmata — result from the inclusion of complementing lemmata that, in contrast to the main paradigmatic arrangement, show a syntagmatic composition as a single textual block. This block, accessed through the main lemma, is composed of the articles associated with each of the subsequent sublemmata (Gouws 2003). Such clusters of sublemmata can be classified into two different categories, niches and nests, depending on whether they adhere to the prevailing alphabetical order (Bergenholtz and Tarp 1995; Gouws 2003; Hausmann and Wiegand 1989; Wiegand 1989).

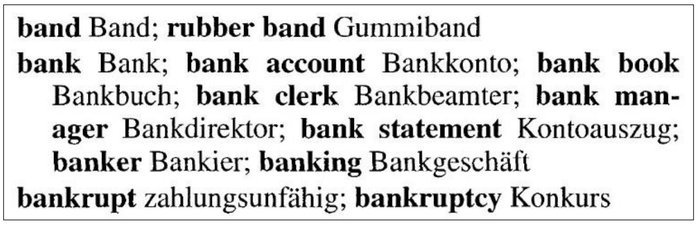
As Figure 1 shows, the sublemmata grouped in niches adhere to an alphabetical order, not only in their internal (horizontal) organization but also concerning the external (vertical) arrangement of the main lemma corpus. In this way, the sublemmata integrated within the niche, in addition to showing an internal alphabetical order, alphabetically precede the following main lemma: "This type of cluster merely illustrates a deviation in the direction of macrostructural ordering, i.e. horizontal instead of vertical, but does not imply any deviation from the prevailing straight alphabetical ordering" (Gouws 2003: 41).



band Band
bank Bank; **bank account** Bankkonto; **bank book** Bankbuch; **bank clerk** Bankbeamter; **bank manager** Bankdirektor; **bank statement** Kontoauszug; **banker** Bankier; **banking** Bankgeschäft
bankrupt zahlungsunfähig; **bankruptcy** Konkurs

Figure 1: Illustrative example of a niche grouping procedure in an English–German bilingual dictionary (Bergenholtz and Tarp 1995)

Nesting differs from niche grouping in one important aspect: although the sublemmata included in the nest follow the preceding main lemma alphabetically, the nest includes sublemmata that do not conform with the alphabetical value of the succeeding main lemma: "[A]s opposed to niching, nesting enables interruption within the nest of the order of graphemes in the access alphabet in order that all lemmata with the same stem may appear together in the same article" (Bergenholtz and Tarp 1995: 194). This deviation from the strict alphabetical order is shared by the two types of nesting: first-level and second-level nesting. As displayed in Figure 2, first-level nesting represents an intermediate stage between the niche and the second-level nesting: the arrangement of sublemmata in this first-level nesting obeys to a strict alphabetical order, however, some sublemmata alter this maxim concerning the following main lemma.



band Band; **rubber band** Gummiband
bank Bank; **bank account** Bankkonto; **bank book** Bankbuch; **bank clerk** Bankbeamter; **bank manager** Bankdirektor; **bank statement** Kontoauszug; **banker** Bankier; **banking** Bankgeschäft
bankrupt zahlungsunfähig; **bankruptcy** Konkurs

Figure 2: Illustrative example of first-level nesting in an English–German bilingual dictionary (Bergenholtz and Tarp 1995)

This strict internal alphabetical arrangement of the first level nesting constraints its lexicographic functionality as compared with the degree of "sophistication" of the second level nesting (Gouws 2003: 41), characterized by a further internal cancellation of the alphabetical order, as will be seen in the WfWG: "[S]econd level nesting gives evidence of a lexicographic procedure where morphosemantic motivations dominate the alphabetical ordering principle in the presentation of sublemmata in a horizontal lemma file" (ibid.: 43).

1.2 The macrostructure of WfWG

The *Wortfamilienwörterbuch der deutschen Gegenwartssprache* is the first and only one-volume learner's dictionary of contemporary German whose macrostructure is organized around word families (Augst 2009).[4] According to Augst (ibid.), this arrangement in word families plays a fundamental role in the acquisition of German (both L1 and L2) to the extent that it facilitates an easy comprehension of the internal mechanisms governing word formation and, by extension, of the more general patterns that connect and organize the lexical units at a higher order level: the grammar.[5]

Augst's proposal (1992: 34) in the WfWG arises from an approach, according to which "in der Wortbildung selbst (wie aber auch in der Wortbildungstheorie) Produktivität auf Grund genereller Regeln und Produktivität auf Grund singulärer Analogie nebeneinander (be)stehen [in the very practice of lexical formation (as well as in the theory of lexical formation) coexist both rule-based and analogy-based productivity; our translation]".[6] On this basis, Augst proposes that, in order to appropriately recreate the lexicon's word-family structure, its lexicographic representation must conform to the "relative motivation" (*relative Motiviertheit*; 2009: IX) between the formal manifestations of the lexical units at a given moment. The starting point of Augst's lexicographic approach lies in the "synchronous etymological competence" (*synchrone etymologische Kompetenz*; 1975: 156-231) understood as any speaker's perceived ability to motivate lexical relations. This ability entails decomposing and reducing the lexicon complexity for the purpose of filtering the morphological core that conveys the central lemma element of a word family (Augst 2009). This scheme, illustrated in Figure 3, facilitates the tracing of an itinerary in the opposite direction towards word formation processes, and the construction, in accordance with the successive derivations (of 1st, 2nd, 3rd degree, etc.), of a hierarchical and recursive or replicating lexical structure (Augst 2009).

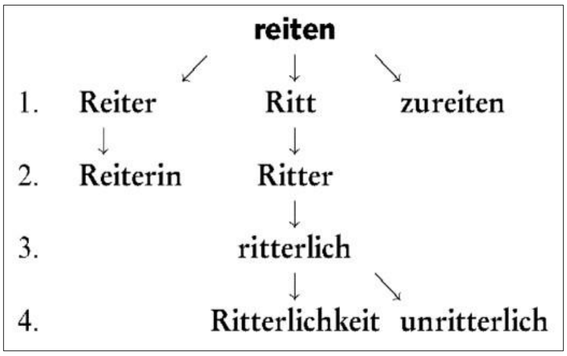
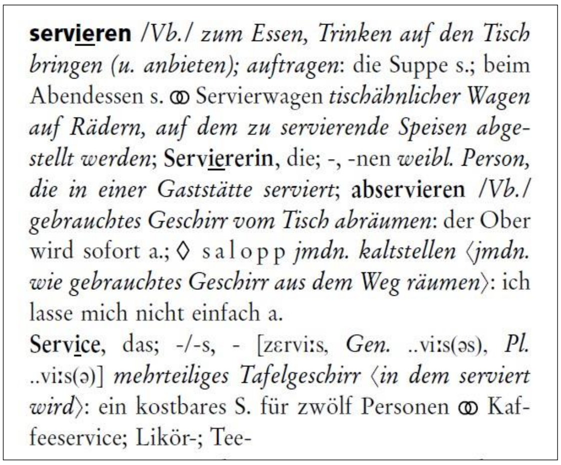


Figure 3: Example of a hierarchical and recursive structure representing the word family *reiten* (Augst 2009)

The arrangement of the lemma corpus according to this type of structure results in a second-level nesting macrostructure (Figure 4). Each headword (*Kernwort*) of the respective word family is listed as a main lemma. Following the headword, the first-degree derivations are listed as sublemmata in independent paragraphs; first, the suffixed derivations, followed by the prefixed derived forms. Second to nth-degree derivations are added horizontally to the first-degree derivations in the corresponding paragraph. The compounds are located after the pertinent sublemma and are identified through the indicator (⚭). The compounds are arranged in such a manner that those compounds in which the sublemma appears as the modifier or *determinans* (*Bestimmungswort*) are listed first. Next, an en-dash precedes compounds in which the sublemma serves as the head or *determinatum* (*Grundwort*). If there is a linking element (*Fugenelement*) between the compound constituents, the sublemmata are sorted alphabetically according to this linking element (for example, compounds with *Wort* have either no linking element or *-er-*, which results in the following arrangement: *Wort* ... ⚭ *Wortart; -gruppe*; *-schatz* — *Wörterbuch* — *Beiwort*). The compounds can, in turn, provide the lexical base for additional second-degree compounds, etc. These compounds appear in parentheses (for example: *Wörterbuch* … (*Bild-*; *Fach-*; *Hand-*).



servieren */Vb./ zum Essen, Trinken auf den Tisch bringen (u. anbieten); auftragen*: die Suppe s.; beim Abendessen s. ⚭ Servierwagen *tischähnlicher Wagen auf Rädern, auf dem zu servierende Speisen abgestellt werden*; **Serviererin**, die; -, -nen *weibl. Person, die in einer Gaststätte serviert*; **abservieren** */Vb./ gebrauchtes Geschirr vom Tisch abräumen*: der Ober wird sofort a.; ◊ salopp *jmdn. kaltstellen* ⟨*jmdn. wie gebrauchtes Geschirr aus dem Weg räumen*⟩: ich lasse mich nicht einfach a.

Service, das; -/-s, - [zɛrviːs, *Gen.* ..viːs(əs), *Pl.* ..viːs(ə)] *mehrteiliges Tafelgeschirr* ⟨*in dem serviert wird*⟩: ein kostbares S. für zwölf Personen ⚭ Kaffeeservice; Likör-; Tee-

Figure 4: Section of the entry related to the lemma *servieren* (Augst 2009) featuring a second-level nesting configuration in the macrostructure

The lexicographic (re)production of this organizing principle reproduces, according to Augst (2009), the structuring of the lexicon around ideal archetypes as perceived by the speaker by virtue of his linguistic competence. In this regard, the conceptual challenge of Augst's lexicographic endeavour lies in "linearizing" the grammatical intricacy of word families into the macrostructure (2009: XII), in such a way that the dictionary user will be able to

(i) recognize the word family hierarchical structure despite the "linearization problem" (Geeraerts 2001: 18) inherent to printed dictionaries, and
(ii) find without difficulty the word that prompted the search, "[d]abei soll die Wortfamilienstruktur für die Bedeutungsangaben jedes einzelnen Wortes der Wortfamilie wechselseitig erhellend wirken und somit die Zerrissenheit des alphabetisch-semasiologischen Wörterbuchs aufheben [yet the word family structure must have a reciprocal highlighting effect on the meaning of each word in the family and thus neutralize the disintegration [of morphosemantic relationships] of the alphabetic-semasiological dictionary; our translation]" (Augst 2009: IX).

2. Material and method

In the following analysis, the macrostructural ordering type is the independent variable whose modification produces an alteration in the corpus AC value as the dependent variable. The test sets will be composed of a relevant lemma collection extracted from the WfWG corpus and subjected to the respective macrostructural arrangements according to (i) the original WfWG arrangement in word families and (ii) the alphabetical arrangement of said corpus. The control set consists of a disordered macrostructure having a random distribution of the same corpus elements.

2.1 Material

The lemma corpus used to perform this study will be extracted from the WfWG (Augst 2009). The lemma selection process to build this corpus will focus on finding "a small and insightful subset" of the original corpus (David et al. 2016). To this end, a double criterion is established: (i) morphological productivity, and (ii) the presence of a lexicographic definition associated with the lemma. According to these parameters, the lemma corpus will consist of the whole set of headwords (including homonyms) and their derived forms. Since word compounding has limited grammatical relevance on a synchronic level, compounds will be excluded. The result generates a set $L = \{l_1, l_2, l_3, \ldots l_{m-1}, l_m\}$ of 27,622 lemmata. This set is treated and presented as a string of n characters belonging to $C = \{c_1, c_2, c_3, \ldots c_{n-1}, c_n\}$, where c_i represents any character $c \in C$ at position i. This procedure results in a string comprising 261,121 characters.

2.2 Method

Algorithmic complexity being formally incomputable, the application of the Minimum Description Length principle (Grünwald 2005, 2007; Rissanen 1978)

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1902 (Article)

will allow us to obtain an estimate of the corpus AC value: "[t]he goal of statistical inference may be cast as trying to find regularity in the data. 'Regularity' may be identified with 'ability to compress.' [Minimum Description Length] combines these two insights by *viewing learning as data compression* [emphasis in the original]" (Grünwald 2007: 12). In other words, the more compressed the data set, the greater the extent to which a system can be said to have learned on that set (Chaitin 2006; Maguire et al. 2015).

In accordance with this methodological framework, the following experimental design is based on the study conducted by Koplenig et al. (2017) on the statistical correspondence between the internal structuring of the lexicon and its syntagmatic ordering. The adopted notations as well as the ensuing explanations derive from said study.

Our interest is focused on determining the amount of regularity of the lemma corpus as a variation of its entropy rate (Koplenig et al. 2017). On this basis, the absolute redundancy (D) will serve as the reference magnitude. This magnitude measures the difference between the absolute rate of entropy (R_0) — that is, its maximum value — and the real or effective rate of entropy (r), being a high value of D indicative of a greater amount of regularity in the set (Koplenig et al. 2017).

$$D = R_0 - r \tag{1}$$

In order to isolate the amount of absolute redundancy (D) that can be attributed to the different lemma arrangements, we will first estimate the entropy value of the control set, that is, of the set L in randomized order. This value corresponds to the absolute entropy rate, R_0, associated to the set L. In order to obtain optimal results randomization will be performed for 1,000 iterations. Subsequently, the set L will be ordered according to the original WfWG word-family arrangement and the effective entropy rate of the resulting set will be estimated. This value will be called r_{WfWG}. The difference, $R_0 - r_{WfWG}$, will give an estimate of the amount of absolute redundancy, D_{WfWG}, that can be attributed to the set L arranged in word families. Finally, the set L will be rearranged in an alphabetical order and its effective entropy rate will be calculated. This value will be called r_{Alpha}. The difference between the absolute rate of entropy, R_0, and the effective entropy rate, r_{Alpha}, will render an approximation to the amount of regularity, as D_{Alpha}, contained in L after imposing an alphabetical order on it.

To obtain the entropy rate value, the non-parametric estimation method developed by Kontoyiannis (1997; Kontoyiannis et al. 1998) will be implemented. This string-match method is closely related to the Lempel–Ziv–Welch compression algorithm (Welch 1984; Ziv and Lempel 1978), where H represents the average amount of entropy estimated at each position i of a string of length N[7]:

$$H = \left[\frac{1}{N} \sum\nolimits_{i=2}^{N} \frac{\ell_i}{\log_2 i}\right]^{-1} \tag{2}$$

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1902 (Article)

The magnitude of interest for the calculation of said amount at each position i is the maximum length of coincidence (maximum string length), ℓ_i. In order to determine the regularity or redundancy at position i, we must first analyze the previous segment of the string up to — but not including — i (Welch 1984), and monitor how many of the characters initials of the segment of the string starting at i have already appeared in the same order somewhere in the previous segment parsed. The value of ℓ_i is obtained by adding the unit to the length of the longest matching substring and thus meeting the following criteria: (i) it starts at position i of the string and (ii) it is not a substring of the string segment before i (Koplenig et al. 2017): "the intuitive idea behind this approach is that longer match-lengths are, on average, indicative of more redundancy in the text and, therefore, a lower mean uncertainty per character" (2017: 3). This amount of entropy contained in each character can be defined as the average amount of information in bits per character (bpc) necessary to reproduce the lemma corpus in the considered macrostructural arrangement (cf. Koplenig et al. 2017).

3. Results

Table 1 presents the results obtained for the different arrangements considered in this study. The WfWG macrostructural arrangement in word families shows a high degree of correlation between the lemma elements and, therefore, a low degree of complexity. Secondly, an alphabetic restructuring at the lexical level results in a disruption of the correlations between the elements, which leads to a higher degree of complexity in the string. Thirdly, a random restructuring removes all regularity contained in the original corpus and results in the highest value of complexity. This value is located in the upper bound for the maximum entropy of the lemma corpus at the lexical level.

Table 1: Sample of the resulting string together with the H value for each experimental setting (ES). The test sets correspond to the settings ES1 and ES2, while the control set is defined in ES3. The table includes the sample standard error (SE) and 95% confidence interval (CI) for the ES3 estimates.

Description	**Sample**	**H Value**	**SE**	**95% CI**
ES1. WfWG	hüllen hülse enthülsen einhüllen enthüllen enthüllung umhüllen verhüllen verhüllt unverhüllt human inhuman	2.31909		
ES2. Alphabetical	einhüllen enthüllen enthüllung enthülsen hüllen hülse human inhuman umhüllen unverhüllt verhüllen verhüllt	2.34744		
ES3. Random	verhüllt enthülsen hülse inhuman einhüllen verhüllen enthüllen unverhüllt umhüllen hüllen enthüllung human	2.40982	2.23733^{-5}	4.38518^{-5}

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1902 (Article)

The difference in the value $D_{WfWG} = .09073$ as compared to $D_{Alpha} = .06238$ indicates a percent increase of 45,44% in regularity for the original word-family arrangement of the WfWG relative to the alphabetical arrangement (Figure 5).

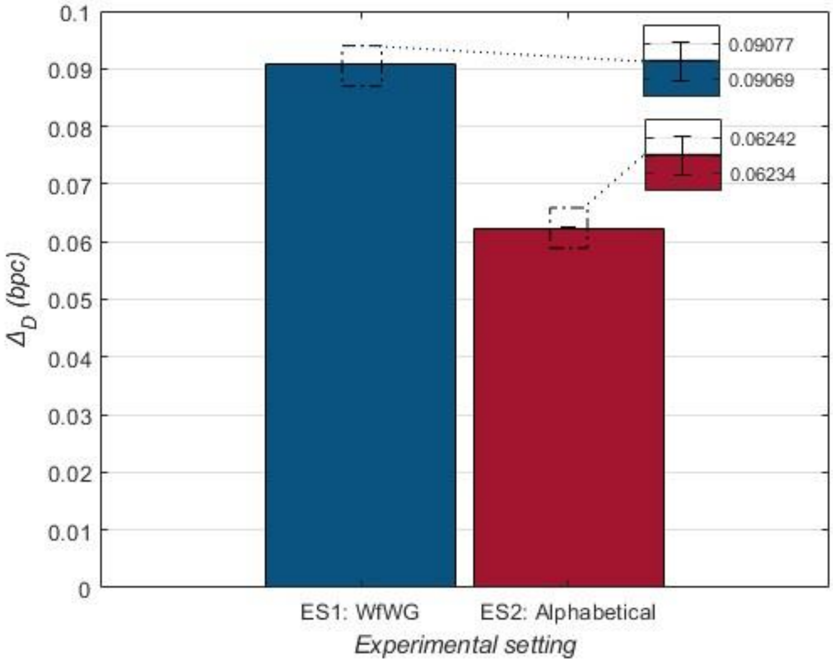


Figure 5: Representation of the increase of D_{WfWG} and D_{Alpha} in bpc according to ES1 (word families) and ES2 (alphabetical) respectively, measured against the control set established in ES3 (random). The error bars represent the 95% confidence interval for 1.000 iterations in ES3.

The increase observed in ES1 and ES2 in relation to ES3 provides a measure of the average amount of regularity or meaningful information (in bpc) gained as a function of the macrostructural arrangement of the lemma set (see Table 2). As Figure 5 shows, this difference corresponds to a percent increase of 39,12% and 26,57%, respectively.

Table 2: Amount of regularity gained by increasing the internal order of the words ($\Delta^{ESx}_{ESx.3}$) as against the amount of regularity gained by the increment in external (macrostructural) order of the words (Δ^{ESx}_{ES3})

Description	Sample	H Value	Δ^{ESx}_{ES3}	$\Delta^{ESx}_{ESx.3}$
ES1. WfWG	hüllen hülse enthülsen einhüllen enthüllen enthüllung umhüllen	2.31909	0.09073 CI: 4.38518 $^{-5}$	1.20267 CI: 4.77847^{-5}
ES3. Random	verhüllt enthülsen hülse inhuman einhüllen verhüllen enthüllen	2.40982		
ES1.3	nlhleü hsüle ehtnnüles lnhilnüee lhlteneün uelnhntglü lülmeuhn	3.52176		
ES2. Alphabetical	einhüllen enthüllen enthüllung enthülsen hüllen hülse human inhuman	2.34744	0.06238 CI: 4.38518 $^{-5}$	1.17259 CI: 4.62447^{-5}
ES3. Random	verhüllt enthülsen hülse inhuman einhüllen verhüllen enthüllen	2.40982		
ES2.3	hlünnelie hüleetnnl üetullnghn lsetnhüne ehünll lsüeh uanhm nmanhui	3.52003		

The values obtained in ES1 and ES2 in relation to ES1.3 and ES2.3 reveal the average gain of redundancy or meaningful information as a function of the intralexical order, that is, of the internal structure of words. This average gain returns values of $\Delta^{ES1}_{ES1.3}$= 1.20267 and $\Delta^{ES2}_{ES2.3}$= 1.17259, respectively. According to these data as presented in Table 3, the intralexical order in the word-family and alphabetical arrangements increases the amount of meaningful information approximately by a multiple of 13 and 17, respectively, as compared to the amount of meaningful information gained from the extralexical order of the lemma corpus.

As for the regularity increase within the alphabetic sections that make up the dictionary, Figure 6 displays the normalized absolute redundancy values (D) relative to each of the sets C_A,C_B,C_C,..., etc., except for the sets C_X and C_Y, whose content (five lemma elements in each section, with a total of 34 and 19 characters, respectively) returns an insufficient amount of data for the application of the proposed method.

In order to validate the data, the constituent elements of the test sets and the control set have been subjected to randomization in subsequent stages (Table 3). This gradual dismantling of the structures progressively suppresses any correlation and, therefore, any regularity present in the corpus.

On the other hand, as Koplenig et al. (2017) observe, the entropy rate H of any process can only be determined in the limit, that is, in strings of infinite length. In this regard, Figure 7 shows that the estimation method presented in (2) yields values that rapidly converge to the entropy source, which suggests that the obtained values yield a valid estimate of said source (ibid.).

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1902 (Article)

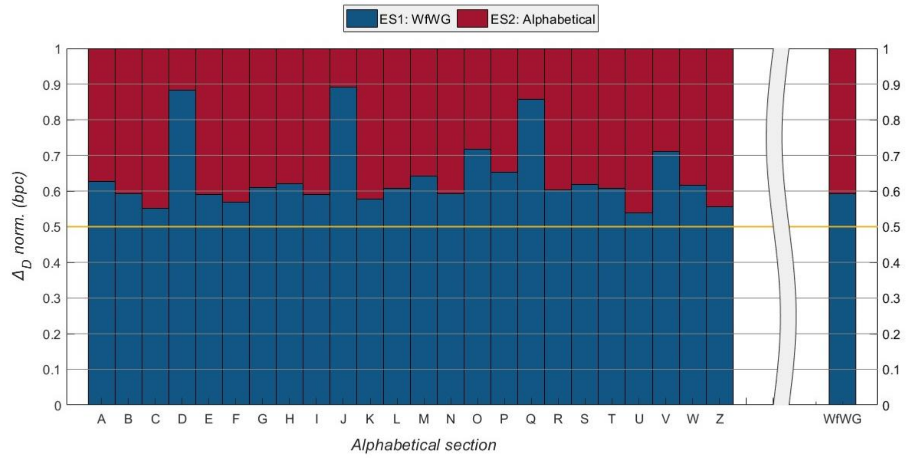


Figure 6: Normalized representation of the increase in bpc of absolute redundancy (D) for each of the alphabetical sections (except for the sets C_X and C_Y) and the whole lemma set (identified as WfWG on the rightmost side of the chart) as a function of the macrostructural arrangement per ES1 (blue) and ES2 (red) compared to the purely random arrangement in ES3. The solid orange line serves as a visual reference indicator illustrating equal values for ES1 and ES2.

Table 3: Dismantling stages for the respective experimental contexts. The third stage — represented by ES1.3, ES2.3, and ES3.3 — reproduces a version of the respective string in which the intralexical regularities have been concealed by replacing the substring containing each lexical element with another substring constructed randomly from the characters available in it (Koplenig et al. 2017). In ES3.5 the spaces have been removed from the string, subsequently the characters have been randomized and the spaces randomly re-inserted in such a way that the corpus extension remains unaltered at 27.622 instances. Results for randomized strings reflect 1,000 iterations.

		Description	Sample	H Value	SE	95% CI
ES1.	**1.1.**	**No spacing**	hüllenhülseenthülseneinhüllenenthüllenenthüllungumhüllenverhüllenverhülltunverhüllthumaninhuman	2.47437		
	1.2.	**Random spacing**	hü ll enhülseenthülseneinhü llenent hüllenenthüllu ngumhülle nve rhül lenverhül ltu nverhüllthumaninhuman	2.78207	3.6103^{-5}	7.07619^{-5}
	1.3.	**Randomized characters with original spacing**	nlhleü hsüle ehtnnüles lnhilnüee lhlteneün uelnhntglü lülmeuhn ervelühnl rlthvlüe nürtlhuvel umhna ahuninm	3.52176	2.43799^{-5}	4.77847^{-5}

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1902 (Article)

ES2.	**2.1. No spacing**	einhüllenenthüllenenthüllungenthülsenhüllenhülsehumaninhumanumhüllenunverhülltverhüllenverhüllt	2.50761		
	2.2. Random spacing	ein hüllenenthü lle nenthüllungenthülsenhüllenhülse humaninhum anumhül l enunv erhülltverhüllenverhül lt	2.77471	3.24735^{-5}	6.36481^{-5}
	2.3. Randomized characters with original spacing	hlünnelie hüleetnnl üetullnghn lsetnhüne ehünll lsüeh uanhm nmanhui nuleühlm üvhrulnetl lülenherv ührtvlle	3.52003	2.35942^{-5}	4.62447^{-5}
ES3.	**3.1. No spacing**	verhülltenthülsenhülseinhumaneinhüllenverhüllenenthüllenunverhülltumhüllenhüllenenthüllunghuman	2.58352	2.95501^{-5}	5.79182^{-5}
	3.2. Random spacing	verhüllte nt hülsenhülse in huma neinhüllenverhüllenent h üllen unverhülltumhüllenhüllen enthüllunghuman	2.82903	3.47808^{-5}	6.81704^{-5}
	3.3. Randomized characters with original spacing	etvhrüll tlennsheü lehüs minanhu leenlihün eelvrünlh htnenülel vtulüehlnr ulhnümle leünhl lüluehtnng anhum	3.53895	2.32421^{-5}	4.55545^{-5}
	3.4. Randomized characters without spacing	onisusnpittoiieoieeaeklnnfpbolheemshrarsielgblenpefiispeetafcüeregsaenclfaiavssoesnaeelmctkrgeorinbrntzkimfue	3.56282	2.19787^{-5}	4.30784^{-5}
	3.5. Randomized characters with random spacing	sgdltaluc tineen ainc rgzatzrsrneb ihbss ntsviikk nnhueur nieguh wbrsedcgmeenkrat ssnakhtek iarseeecoe äeheudzane	3.571009	2.06591^{-5}	4.04917^{-5}

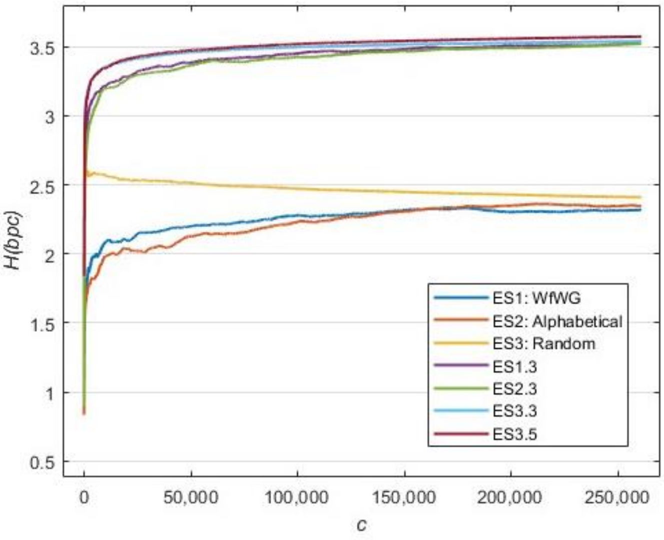


Figure 7: Entropy rate H in bpc as a function of incorporated lemmata expressed as c. The results attest that a small amount of data is enough to demonstrate a convergence towards the entropy source (Koplenig et al. 2017). All three experimental settings ESx.3 (randomized characters with original spacing) deal with the internal order

of words in their respective macrostructural arrangement. In ES3.5 (randomized characters with random spacing), after removing the spaces from the string, the characters have been randomized and the spaces randomly re-inserted into the string in such a way that the total amount of 'words' remains unaltered at 27.622 instances.

4. Discussion

The specific objectives of this study have focused on (O1) determining whether the macrostructural arrangement has an impact on the AC value of the WfWG corpus, (O2) calculating the difference between the corpus macrostructural arrangements — in word families and alphabetical — with regard to their AC values, as well as (O3) identifying the macrostructural arrangement that delivers the less complex version of said corpus. For this purpose, the Minimum Description Length principle has been applied to estimate the corpus AC value based on its compressibility. Among the findings, with reference to O1 the fundamentals in the data show that the macrostructural arrangement influences the AC value of the WfWG corpus. Concerning O2, the data also reveal that alterations in the macrostructure towards more ordered arrangements significantly decrease the AC value of the corpus. However, the most significant finding of this study, in relation to O3, reveals that the WfWG corpus in a word-family arrangement has a lower AC value in comparison to the value associated with the same corpus arranged in alphabetical order (see Table 1). As observed from Figure 5, said relative minimum value represents a significant increase in regularity for the word-family distribution, approximately doubling the regularity gain of the alphabetical layout relative to a purely random arrangement. On the other hand, the estimates for each alphabetic section of the dictionary displayed in Figure 6 manifest that, as for the entire set, the arrangement in word families contributes to lower AC values. The fluctuation in the relative values across the alphabetic sections points to a variation in the numerical proportion of word families in relation to the number of those lemmata that do not appear attached to any of the listed word families. Furthermore, in light of the validation procedures applied, the consistency of the results suggests that the Minimum Description Length principle renders a valid method to estimate the complexity associated with the corpus macrostructural arrangement. In more general terms, an AIT-related linguistic interpretation of the results allows us to argue that analogy is language's algorithmically simpler and, therefore, more efficient operation mode, driving it — as a self-regulating natural system — towards a more ordered configuration distant from entropic degradation (Devine 2020). In this sense, the present study inscribes itself in the area of research dedicated to the widely analyzed and documented phenomenon of analogy as a fundamental cognitive strategy in language processing.

Notwithstanding the coherence of the results with respect to the guiding principles of our proposal, this approach reveals certain limitations. An inherent drawback resides in the restriction of the corpus to a set of lexical items whose superior demarcation is the word. This excludes the consideration of (to a greater or lesser degree) lexicalized syntactic constructions, for example, constructions with a functional verb, or *Funktionsverbgefüge*, in which the grammatical function of the construct is expressed by the verb. Additionally, and due to the purely numerical nature of this approach, the results do not allow us to present explicit-declarative statements about the grammar — which as a description of minimum length (Kornai 2008) governs the corpus — nor about the nature of the structures or patterns affected by the compression (both at the intralexical and interlexical level). An observation that, on the other hand, suggests that grammar acquisition through inductive inference is related to the formation of (primarily) implicit knowledge in procedural memory (DeKeyser 2015; Paradis 2009; Ullman 2016). In addition, constraining the arrangement criteria of the test sets to both word-family and alphabetical principles leaves semantic-oriented arrangement criteria unconsidered. In this regard, since cognitive-semantic criteria overlook the formal component of the linguistic sign, the macrostructural arrangements conforming to purely semantic precepts shall be deemed random per the approach adopted in this research, with their AC estimates expected to surpass the AC numbers for word-family and alphabetical arrangements, leaning towards maximum entropy values. However, the most important limitation of this study lies in the reduction of the object of analysis to a single language with distinct typological characteristics. If, as the study by Koplenig et al. (2017) suggests, the entropy rate of any language lemma corpus is determined by morphological factors, we believe that future studies including the morphological typology as an independent variable can introduce additional evidence that would validate the approach pursued in this study.

In contrast to the prevalent approaches in monolingual learner lexicography based on theoretical frameworks of cognitive semantics, our proposal introduces a psycho-cognitive approach based on a quantifiable, irreducible, and theory-neutral notion of complexity and, by extension, of learnability. In this regard, although the macrostructural arrangement in word families has been widely applied and referred to in the German lexicographic tradition as a method to facilitate language learning (by exposing the internal mechanisms of lexical production), these results suggest that the methodology adopted in this study would enable, for the first time, a theory-neutral quantitative evaluation of the didactic nature of this lexicographic practice.

On a separate note, this proposal based on the complexity-learnability binomial also supports the idea of implementing a differentiated approach in the area of digital lexicography (Bothma 2011, 2017; Bothma et al. 2016) whose interest would be focused, beyond optimal searching, in the navigational (browsing) processing of the lemma set. According to the psycho-cognitive and organizational foundations of Neuroergonomics (Lapeyre et al. 2011; Li and Klippel 2016; Montello 2005; Montello and Sas 2006), navigation in less complex (irrespective of the formal definition of complexity applied) and thus more regular environments promotes spatial awareness and the creation of more complete and

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1902 (Article)

precise cognitive maps that facilitate orientation and, thereby, the understanding of the environment or search space considered, in our case, the lemma corpus.

Finally, we consider it necessary to recapitulate that the current study does not address the quantitative determination of the corpus learnability. In this respect, an empirical analysis that would provide a quantitative measure of the corpus learnability as a function of the AC value associated to its macrostructural arrangement represents the major research objective of future studies within the framework of this project.

5. Conclusions

The general objective of this study is to promote language acquisition in monolingual learner's dictionary users from a psycho-cognitive perspective. In our proposal, we view the human cognition essentially as a learner sensitive to fluctuations in the environmental complexity in such a way that our learning efficiency increases in less complex environments (Kempe et al. 2015; Zenil and Gauvrit 2017). Against this background, our methodological approach — the Minimum Description Length principle — allowed us to obtain a quantitative estimation of the algorithmically bound complexity (AC) attributed to the WfWG lemma corpus according to its macrostructural arrangement. The resulting data indicate that, compared to the alphabetical layout, the arrangement in word families provides a more ordered, less complex, and, by extension, more learnable version of the lemma corpus. These results open a door to future studies with the aim to determine the lemma corpus variation in learnability as a function of the AC value derived from its macrostructural arrangement. We hope that this psycho-cognitive approach based upon the principles and practical methods of AIT may well be useful in implementing macrostructural designs in monolingual learner lexicography (both paper and digital) which would reinforce language acquisition in the dictionary user.

Acknowledgements

Our deep appreciation and gratitude are due to those whose inestimable peer review and editorial guidance vastly enhanced our original draft. Any remaining errors, omissions, and limitations are entirely ours.

Endnotes

1. De Schryver (2013) points out that in computational lexicography, analogous to printed lexicography, the macrostructural design refers not only to the dictionary as an ordered arrangement of the lemma corpus but mainly to the set of interlexical relationships that are configured around parameters such as grammatical category, morphology, valence, semantic features, etc.
2. The macrostructural treatment in machine-readable dictionaries of this particular type tends to reflect a methodical and consistent digitalization of the same theoretical principles and, beyond

the formal qualities of the new medium (menus, hypertexts, multimedia, etc.), they do not differ substantially from their printed counterparts (Chen 2012; Dziemianko 2017; Kobayashi 2007).

3. This approach represents a concretion in the area of monolingual learner lexicography of the "simplicity principle" applied in the field of SLA (Chater and Vitányi 2007; Chater et al. 2015). According to this principle, "the learner has sufficient data to learn successfully from positive evidence if it favors the simplest encoding of the linguistic input [emphasis in the original]" (Hsu et al. 2013: 35).
4. A didactically motivated approach to the word-family arrangement in German lexicographic tradition dates back as far as 1700 with the publication of *Das herrlich grosse teutsch–italiänische Dictionarium* by Matthias Kramer (cf. Haß-Zumkehr 2012: 81-88). Kramer relates his decision in favor of a word-family arrangement to the practical didactic requirements for the production of new words, that is, to the apprehension of the language internal (grammatical) mechanisms (ibid.: 84-85). Hence, the macrostructural layout in word families of his dictionary is essentially justified by its didactic purpose, an attribute that, in Kramer's words, prevails over its functionality as a reference work (Wiegand 1998: 657).
5. Consistent with Cruse's definition, a "lexical unit" designates "the union of a lexical form and a single sense" (1986: 77) as opposed to a "lexeme", a term that refers to a cluster of lexical units: "lexemes, on the other hand, are the items listed in the lexicon, or 'ideal dictionary', of a language" (Cruse 1986: 49). The lexeme congregates a group of lexical units sharing a common root and therefore, maintaining a certain relationship, both in their phonetic composition and in their meaning (Bergenholtz and Tarp 1995; Umbreit 2011): "[A] dictionary contains (among other things) an alphabetical list of the lexemes of a language. We shall characterise a lexeme as a family of lexical units" (Cruse 1986: 76). A "lexical item", on the other hand, designates "any word, abbreviation, partial word, or phrase which can figure in a dictionary (often as the headword of an entry) as the 'target' of some form of lexicographic description, most commonly a definition or a translation" (Atkins and Rundell 2008: 163).
6. Augst points here to the primary, and often fuzzy distinction between analogy and rule. In this regard, Kiparsky (1975, in Derwing and Skousen 1989: 56) argues that it is problematic to draw a clear, rigorous, and unequivocal boundary between both, since "at the point at which [...] analogies begin to make the right generalizations, they are indistinguishable from rules." Along the same lines, Haspelmath (2002: 103) considers morphological analogy and regularity as "really one and the same thing", while in the words of Krott (2009: 118) rules can be qualified "as extreme case of analogy".
7. Koplenig et al. (2017) remark that, since this is an estimate of the theoretical (and unobservable) value, the correct mathematical notation corresponds to the hat notation $\hat{H}$. In this regard, we adhere to the authors' motivation and, for the sake of simplicity, we adopt the plain notation H.

References

A. Dictionaries

Augst, Gerhard. 2009. *Wortfamilienwörterbuch der deutschen Gegenwartssprache*. Tübingen: Niemeyer. https://doi.org/10.1515/9783484971332

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1902 (Article)

Benson, Morton, Evelyn Benson and Robert Ilson. 2009. *The BBI Combinatory Dictionary of English. Your Guide to Collocations and Grammar*. Amsterdam: John Benjamins.

Bosque, Ignacio (Ed.). 2006. *Diccionario combinatorio práctico del español contemporáneo. Las palabras in su contexto.* Madrid: SM.

Casares, Julio. 2013. *Diccionario ideológico de la lengua española. Desde la idea a la palabra, desde la palabra a la idea.* Barcelona: Gustavo Gili.

Davau, Maurice, Marcel Cohen and Maurice Lallemand. 1984. *Dictionnaire du français vivant*. Paris: Bordas.

Dornseiff, Franz. 2020. *Der deutsche Wortschatz nach Sachgruppen*. Berlin/Boston: De Gruyter. https://doi.org/10.1515/9783110457742

Häcki Buhofer, Annelies, Marcel Dräger, Stefanie Meier and Tobias Roth. 2014. *Feste Wortverbindungen des Deutschen. Kollokationenwörterbuch für den Alltag*. Tübingen: Francke.

Kirkpatrick, Elizabeth M. (Ed.). 1983. *Chambers Universal Learners' Dictionary*. Edinburgh: Chambers.

Mel'čuk, Igor A., Nadia Arbatchewsky-Jumarie, Lidija Iordanskaja, Suzanne Mantha and Alain Polguère. 1999. *Dictionnaire explicatif et combinatoire du français contemporain*. Montreal: Montreal University Press.

Rey-Debove, Josette and Alain Rey. 2009. *Le nouveau Petit Robert. Dictionnaire alphabétique et analogique de la langue française.* Paris: Dictionnaires Le Robert.

Simone, Raffaele. 2010. *Grande Dizionario Analogico della Lingua Italiana.* Turin: UTET.

B. Other Literature

Atkins, B.T. Sue and Michael Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford/New York: Oxford University Press.

Augst, Gerhard. 1975. *Untersuchungen zum Morpheminventar der deutschen Gegenwartssprache.* Tübingen: Narr.

Augst, Gerhard. 1992. Das lexikologische Phänomen der Wortfamilie in alphabetisch-semasiologischen Wörterbüchern. *Zeitschrift für germanistische Linguistik* 20(1): 24-36. https://doi.org/10.1515/zfgl.1992.20.1.24

Bergenholtz, Henning and Sven Tarp (Eds.). 1995. *Manual of Specialised Lexicography. The Preparation of Specialised Dictionaries.* Amsterdam: John Benjamins.

Bothma, Theo J.D. 2011. Filtering and Adapting Data and Information in an Online Environment in Response to User Needs. Fuertes-Olivera, P.A. and H. Bergenholtz (Eds.). 2011. *e-Lexicography. The Internet, Digital Initiatives and Lexicography*: 71-102. London: Bloomsbury Academic.

Bothma, Theo J.D. 2017. Lexicography and Information Science. Fuertes-Olivera, P.A. (Ed.). 2017. *The Routledge Handbook of Lexicography*: 197-216. London: Routledge. https://doi.org/10.4324/9781315104942-14

Bothma, Theo J.D., Rufus H. Gouws and Danie J. Prinsloo. 2016. The Role of e-Lexicography in the Confirmation of Lexicography as an Independent and Multidisciplinary Field. Margalitadze, T. and G. Meladze (Eds.). 2016. *Proceedings of the 17th EURALEX International Congress. Lexicography and Linguistic Diversity, Ivane Javakhishvili Tbilisi State University, Tbilisi, Georgia, 6–10 September 2016*: 109-116. Tbilisi: Ivane Javakhishvili Tbilisi State University.

Chaitin, Gregory J. 1969. On the Length of Programs for Computing Finite Binary Sequences. Statistical Considerations. *Journal of the ACM* 16(1): 145-159. https://doi.org/10.1145/321495.321506

Chaitin, Gregory J. 2004. *Algorithmic Information Theory*. Cambridge, UK: Cambridge University Press. https://doi.org/10.1017/CBO9780511608858

Chaitin, Gregory J. 2006. The Limits of Reason. *Scientific American* 294(3): 74-81.

Chater, Nick, Alexander Clark, John A. Goldsmith and Amy Perfors. 2015. *Empiricism and Language Learnability*. Oxford, UK: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198734260.001.0001

Chater, Nick and Paul Vitányi. 2007. 'Ideal Learning' of Natural Language. Positive Results about Learning from Positive Evidence. *Journal of Mathematical Psychology* 51(3): 135-163. https://doi.org/10.1016/j.jmp.2006.10.002

Chen, Yuzhen. 2012. Dictionary Use and Vocabulary Learning in the Context of Reading. *International Journal of Lexicography* 25(2): 216-247. https://doi.org/10.1093/ijl/ecr031

Clark, Alexander and Shalom Lappin. 2013. Complexity in Language Acquisition. *Topics in Cognitive Science* 5(1): 89-110. https://doi.org/10.1111/tops.12001

Cruse, David A. 1986. *Lexical Semantics*. Cambridge, UK: Cambridge University Press.

David, Ofir, Shay Moran and Amir Yehudayoff. 2016. Supervised Learning through the Lens of Compression. Lee, D. et al. (Eds.). 2016. *Advances in Neural Information Processing Systems 29. 30th Conference on Neural Information Processing Systems, Barcelona, 5–10 December 2016:* 2784-2792.

DeKeyser, Robert. 2015. Skill Acquisition Theory. Van Patten B. and J. Williams (Eds.). 2015. *Theories in Second Language Acquisition*: 94-112. Second edition. New York: Routledge.

Derwing, Bruce L. and Royal Skousen. 1989. Morphology in the Mental Lexicon. A New Look at Analogy. Booij, G. and J. van Marle (Eds.). 1989. *Yearbook of Morphology 1989. Volume 2*: 55-71. Dordrecht: Kluwer Academics. https://doi.org/10.1515/9783112420560-005

De Schryver, Gilles-Maurice. 2013. Tools to Support the Design of a Macrostructure. Gouws, R.H., U. Heid, W. Schweickard and H.E. Wiegand (Eds.). 2013. *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography:* 1384-1395. Berlin/Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110238136.1384

Devine, Sean. 2020. *Algorithmic Information Theory for Physicists and Natural Scientists*. Bristol: IOP Publishing. https://doi.org/10.1088/978-0-7503-2640-7

Dong, Zhendong and Qiang Dong. 2006. *HowNet and the Computation of Meaning*. Singapore: World Scientific. https://doi.org/10.1142/5935

Dziemianko, Anna. 2017. Dictionary Form in Decoding, Encoding and Retention: Further Insights. *ReCALL* 29(3): 335-356. https://doi.org/10.1017/S0958344017000131

Engelberg, Stefan and Lothar Lemnitzer. 2009. *Lexikographie und Wörterbuchbenutzung*. Tübingen: Stauffenburg.

Evans, Vyvyan. 2009. *How Words Mean. Lexical Concepts, Cognitive Models, and Meaning Construction*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199234660.001.0001

Fillmore, Charles J. 1982. Frame Semantics. *Linguistics in the Morning Calm. Selected Papers from SICOL-1981*: 111-137. Seoul: Hanshin.

Fillmore, Charles J. 1985. Frames and the Semantics of Understanding. *Quaderni di Semantica* 6(2): 222-254.

Fillmore, Charles J., Christopher R. Johnson and Miriam R.L. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography* 16(3): 235-250.
https://doi.org/10.1093/ijl/16.3.235

Fuertes-Olivera, Pedro A. and Sven Tarp. 2011. Lexicography for the Third Millennium. Cognitive-oriented Specialised Dictionaries for Learners. *Ibérica* 21: 141-161.
https://www.revistaiberica.org/index.php/iberica/article/view/332

Fulop, Sean A. and Nick Chater. 2013. Learnability Theory. *WIREs Cognitive Science* 4(3): 299-306.
https://doi.org/10.1002/wcs.1228

Geeraerts, Dirk. 2001. The Definitional Practice of Dictionaries and the Cognitive Semantic Conception of Polysemy. *Lexicographica* 17: 6-21.
https://doi.org/10.1515/9783110244212.6

Gouws, Rufus H. 2003. Types of Articles, their Structure and Different Types of Lemmata. Van Sterkenburg, P. (Ed.). 2003. *A Practical Guide to Lexicography*: 34-43. Amsterdam: John Benjamins.

Grünwald, Peter. 2005. Introducing the Minimum Description Length Principle. Grünwald, P., J.I. Myung and M.A. Pitt (Eds.). 2005. *Advances in Minimum Description Length. Theory and Applications*: 3-22. Cambridge, MA: The MIT Press.
https://doi.org/10.7551/mitpress/1114.003.0004

Grünwald, Peter. 2007. *The Minimum Description Length Principle*. Cambridge, MA: The MIT Press.

Grünwald, Peter and Paul Vitányi. 2008. Algorithmic Information Theory. Adriaans, P. and J. van Benthem (Eds.). 2008. *Philosophy of Information*: 289-325. Amsterdam: Elsevier.
https://doi.org/10.48550/arXiv.0809.2754

Haspelmath, Martin. 2002. *Understanding Morphology*. London: Hodder Education.

Haß-Zumkehr, Ulrike. 2012. *Deutsche Wörterbücher: Brennpunkt von Sprach- und Kulturgeschichte*. Berlin/Boston: De Gruyter.
https://doi.org/10.1515/9783110849189

Hausmann, Franz Josef and Herbert Ernst Wiegand. 1989. Component Parts and Structures of General Monolingual Dictionaries. A Survey. Hausmann, F.J., O. Reichmann, H.E. Wiegand and L. Zgusta (Eds.). 1989. *Wörterbücher. Ein internationales Handbuch zur Lexikographie. Volume 1:* 328-360. Berlin: De Gruyter.

Hsu, Anne S., Nick Chater and Paul Vitányi. 2013. Language Learning from Positive Evidence, Reconsidered. A Simplicity-based Approach. *Topics in Cognitive Science* 5(1): 35-55.
https://doi.org/10.1111/tops.12005

Kempe, Vera, Nicolas Gauvrit and Douglas Forsyth. 2015. Structure Emerges Faster during Cultural Transmission in Children than in Adults. *Cognition* 136: 247-254.
https://doi.org/10.1016/j.cognition.2014.11.038

Kiparsky, Paul. 1975. What Are Phonological Theories About? Cohen, D. and J.R. Wirth (Eds.). 1975. *Testing Linguistic Hypotheses*: 187-209. Washington: Wiley.

Kobayashi, Chiho. 2007. Comparing Electronic and Printed Dictionaries: Their Effects on Lexical Processing Strategy Use, Word Retention, and Reading Comprehension. Bradford-Watts, K. (Ed.). 2007. *JALT 2006 Conference Proceedings*: 657-671. Tokyo: Japan Association of Language Teaching.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1902 (Article)

Kolmogorov, Andrei N. 1963. On Tables of Random Numbers. *The Indian Journal of Statistics*. Series A 25(4): 369-376.

Kontoyiannis, Ioannis. 1997. The Complexity and Entropy of Literary Styles. *NSF Technical Report* 97: 1-15.
https://purl.stanford.edu/nw057vj8228

Kontoyiannis, Ioannis, Paul H. Algoet, Yuri M. Suhov and Abraham J. Wyner. 1998. Nonparametric Entropy Estimation for Stationary Processes and Random Fields, with Applications to English Text. *IEEE Transactions on Information Theory* 44(3): 1319-1327.
https://doi.org/10.1109/18.669425

Koplenig, Alexander, Peter Meyer, Sascha Wolfer, Carolin Müller-Spitzer and Kenny Smith. 2017. The Statistical Trade-off between Word Order and Word Structure. Large-scale Evidence for the Principle of Least Effort. *PLoS ONE* 12(3): 1-25.
https://doi:10.1371/journal.pone.0173614

Kornai, András. 2008. *Mathematical Linguistics*. London: Springer.
https://doi.org/10.1007/978-1-84628-986-6

Kövecses, Zoltán and Szilvia Csábi. 2014. Lexicography and Cognitive Linguistics. *Revista Española de Lingüística Aplicada/Spanish Journal of Applied Linguistics* 27(1): 118-139.
https://doi.org/10.1075/resla.27.1.05kov

Kremer, Gerhard, Andrea Abel and Marco Baroni. 2008. Cognitively Salient Relations for Multilingual Lexicography. Zock, M. and C.-R. Huang (Eds.). 2008. *Proceedings of the Workshop on Cognitive Aspects of the Lexicon (COGALEX 2008), Manchester, 24 August 2008:* 94-101. Coling 2008 Organizing Committee. https://aclanthology.org/W08-1913

Krott, Andrea. 2009. The Role of Analogy for Compound Words. Blevins, J.P. and J. Blevins (Eds.). 2009. *Analogy in Grammar. Form and Acquisition*: 118-136. Oxford: Oxford University Press.
https://doi.org/10.1093/acprof:oso/9780199547548.003.0006

Lakoff, George. 1993. The Contemporary Theory of Metaphor. Ortony, A. (Ed.). 1993. *Metaphor and Thought*: 202-251. Cambridge, UK: Cambridge University Press.
https://doi.org/10.1017/CBO9781139173865.013

Lakoff, George and Mark Johnson. 1980. *Metaphors We Live By*. Chicago: University of Chicago Press.

Lapeyre, Brigitte, Sylvain Hourlier, Xavier Servantie, Bernard N'Kaoua and Hélène Sauzéon. 2011. Using the Landmark–Route–Survey Framework to Evaluate Spatial Knowledge Obtained from Synthetic Vision Systems. *Human Factors: The Journal of the Human Factors and Ergonomics Society* 53(6): 647-661.
https://doi.org/10.1177/0018720811421171

Li, Rui and Alexander Klippel. 2016. Wayfinding Behaviors in Complex Buildings. *Environment and Behavior* 48(3): 482-510.
https://doi.org/10.1177/0013916514550243

Maguire, Phil, Oisín Mulhall, Rebecca Maguire and Jessica Taylor. 2015. Compressionism: A Theory of Mind Based on Data Compression. Airenti, G., B.G. Bara and G. Sandini (Eds.). 2015. *Proceedings of the EuroAsianPacific Joint Conference on Cognitive Science / 4th European Conference on Cognitive Science / 11th International Conference on Cognitive Science, Torino, Italy, 25–27 September 2015*: 294-299. CEUR Workshop Proceedings.

Martínez de Sousa, José. 2009. *Manual básico de Lexicografía*. Gijón: Ediciones Trea.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1902 (Article)

Miller, George A., Richard Beckwith, Christiane Fellbaum, Derek Gross and Katherine J. Miller. 1990. Introduction to WordNet. An On-line Lexical Database. *International Journal of Lexicography* 3(4): 235-244.
https://doi.org/10.1093/ijl/3.4.235

Montello, Daniel R. 2005. Navigation. Shah, P. and A. Miyake (Eds.). 2005. *The Cambridge Handbook of Visuospatial Thinking*: 257-294. Cambridge, UK: Cambridge University Press.
https://doi.org/10.1017/CBO9780511610448.008

Montello, Daniel R. and Corina Sas. 2006. Human Factors of Wayfinding in Navigation. Karwowski, W. (Ed.). 2006. *International Encyclopedia of Ergonomics and Human Factors*: 2003-2008. Boca Raton, FL: CRC Press.
http://dx.doi.org/10.1201/9780849375477.ch394

Ostermann, Carolin. 2015. *Cognitive Lexicography: A New Approach to Lexicography Making Use of Cognitive Semantics*. Berlin/Boston: De Gruyter.
https://doi.org/10.1515/9783110424164

Paradis, Michel. 2009. *Declarative and Procedural Determinants of Second Languages.* Philadelphia: John Benjamins.

Richardson, Stephen D., William B. Dolan and Lucy Vanderwende. 1998. MindNet: Acquiring and Structuring Semantic Information from Text. *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics (ACL–Coling), Montreal, Quebec, Canada, August 1998. Volume 2:* 1098-1102. Montreal: Association for Computational Linguistics.
https://dl.acm.org/doi/10.3115/980691.980749

Rissanen, Jorma. 1978. Modeling by Shortest Data Description. *Automatica* 14(5): 465-471.
https://doi.org/10.1016/0005-1098(78)90005-5

Solomonoff, Ray J. 1964a. A Formal Theory of Inductive Inference. Part I. *Information and Control* 7(1): 1-22.
https://doi.org/10.1016/S0019-9958(64)90223-2

Solomonoff, Ray J. 1964b. A Formal Theory of Inductive Inference. Part II. *Information and Control* 7(2): 224-254.
https://doi.org/10.1016/S0019-9958(64)90131-7

Tyler, Andrea and Vyvyan Evans. 2004. Applying Cognitive Linguistics to Pedagogical Grammar: The Case of *Over*. Achard, M. and S. Niemeier (Eds.). 2004. *Cognitive Linguistics, Second Language Acquisition, and Foreign Language Teaching*: 257-280. Berlin/New York: De Gruyter Mouton.
https://doi.org/10.1515/9783110199857.257

Ullman, Michael T. 2016. The Declarative/Procedural Model: A Neurobiological Model of Language Learning, Knowledge, and Use. Hickok, G. and S.L. Small (Eds.). 2016. *Neurobiology of Language*: 953-968. London: Academy Press.
https://doi.org/10.1016/B978-0-12-407794-2.00076-6

Umbreit, Birgit. 2011. Motivational Networks: An Empirically Supported Cognitive Phenomenon. Panther, K.-U. and G. Radden (Eds.). 2011. *Motivation in Grammar and the Lexicon*: 269-286. Amsterdam: John Benjamins.
https://doi.org/10.1075/hcp.27.17umb

Welch, Terry A. 1984. A Technique for High-performance Data Compression. *Computer* 17(6): 8-19.
https://doi.org/10.1109/MC.1984.1659158

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1902 (Article)

Wiegand, Herbert Ernst. 1989. Aspekte der Makrostruktur im allgemeinen einsprachigen Wörterbuch. Alphabetische Anordnungsformen und ihre Probleme. Hausmann, F.J., O. Reichmann, H.E. Wiegand and L. Zgusta (Eds.). 1989. *Wörterbücher. Ein internationales Handbuch zur Lexikographie. Volume 1:* 371-409. Berlin: De Gruyter.
http://dx.doi.org/10.1515/9783110095852.1.4.328

Wiegand, Herbert Ernst. 1998. Historische Lexikographie. Besch, W., A. Betten, O. Reichmann and S. Sonderegger (Eds.). 1998. *Sprachgeschichte. Ein Handbuch zur Geschichte der deutschen Sprache und ihrer Erforschung. Volume 1:* 643-713. Berlin/New York: De Gruyter.

Zenil, Hector and Nicolas Gauvrit. 2017. Algorithmic Cognition and the Computational Nature of the Mind. Meyers, R.A. (Ed.). 2017. *Encyclopedia of Complexity and Systems Science*: 1-9. New York: Springer.
https://doi.org/10.1007/978-3-642-27737-5_707-2

Ziv, Jacob and Abraham Lempel. 1978. Compression of Individual Sequences via Variable-rate Coding. *IEEE Transactions on Information Theory* 24(5): 530-536.
https://doi.org/10.1109/TIT.1978.1055934

Hou Min. *A Dictionary of Chinese Neologisms (2000–2020)*. 2023, XL + 786 pp. ISBN 978-7-100-21777-4 (Hardback). Beijing: The Commercial Press. Price $17.69.

The Chinese language has been developing at a faster pace than ever before, which can be attributed to China's fast economic development and rapid technological advances. As a result, the Chinese vocabulary has experienced exponential growth with new lexical items cropping up at an unprecedented rate. Every year witnesses the addition of at least 800 new Chinese words (excluding many buzzwords, topical words and nonce words), which is comparable to that in the English language. The publication of dozens of dictionaries of Chinese neologisms is reflective of the rapid growth of the Chinese vocabulary. Since the beginning of the 21st century, more than thirty dictionaries devoted to the chronicling of Chinese neologisms have been published.

As one of the most famous researchers on Chinese neologisms, Professor Hou Min has published several neologisms dictionaries in a series entitled *An Annual List of Chinese Neologisms* (汉语新词语). The annual list was first published in 2006 under the editorship of Zhou Jian and Hou Min picked up the baton in 2007. From 2006 to 2018, every year saw the publication of a new edition, each of which records about 300 to 500 neologisms. Its 2015 edition, for instance, includes 471 neologisms that were culled from a corpus of 1.2 billion words, as is exemplified by headwords like 点赞贴 (rave post), 毒丸计划 (poison pill), 孤儿药 (orphan drug), 零工经济 (gig economy), 人口悬崖 (demographic cliff), 私播客 (SPOC), etc. Hou Min passed the baton to Zou Yu, her long-time co-editor in 2018. Since then, a new edition of the list has been published every two years.

In 2023, Hou Min published *A Dictionary of Chinese Neologisms (2000–2020)* which, as is explained in the guide to the use of the dictionary, includes about 4 200 neologisms that were created from 2000 to 2020. As the series upon which DCN was based were compiled with the help of a 5-billion-character corpus, DCN was able to indicate the frequency of its headwords by using a four-star labelling system. Those words appearing over 2 000 times are labelled with five stars, such as 表情包 (meme), 差评 (negative comment), 创客 (creator, maker), 发帖 (to post), 华丽转身 (great makeover) and 僵尸企业 (zombie company). Those appearing less than 100 times are considered low-frequency, such as 鼻影 (nose shadow), 床东 (bedlord college student who rents out his dormitory bed), 导览器 (audio guide), 婚闹 (wedding hazing), 盲约 (blind date) and 情侣衫 (matching couple tops). Another advantage of being corpus-based is the easy extraction of illustrative examples. DCN should be commended for its provision of two examples for almost all of the headwords. 边会 (bilateral or multilateral meeting), for instance, is furnished with two illustrative examples taken from *People's Daily* and *Beijing Evening Paper* respectively.

As DCN takes a corpus-based approach to new-word lexicography, it is able to reflect the latest lexical changes, which can be manifested in its inclu-

Lexikos 34 (AFRILEX-reeks/series 34: 2024): 189-197

sion of two types of words. The first type concerns itself with the COVID-19 pandemic, and DCN has included at least a dozen COVID-related words, such as 鼻拭子 (nose swab), 额温枪 (forehead thermometer), 方舱医院 (mobile hospital), 封城 (lockdown), 健康码 (health code), 流调 (epidemiological survey) and 新冠肺炎 (COVID-19). The second type is related to online words and expressions which cropped up since the beginning of the century. DCN recorded scores of such words and expressions and provided the label "〈网〉" (online) for them. Some of the most popularly used online expressions include 标题党 (one who posts messages and articles using exaggerated or sensationalized titles), 菜鸟 (newbie), 点赞 (to like), 杠精 (person who argues for the sake of arguing, contrarian), 青蛙 (literally a frog, a very ugly man), 社畜 (overworked and exploited employee), 实锤 (smoking gun), 小白 (literally small white, newbie) and so forth.

Etymologically speaking, there are several types of Chinese neologisms. Yang and Yang (2009: 97-98) identified six types, namely new words, new meanings of existing words, dialectal or regional words that are enjoying wider currency, loanwords of all kinds, English initialisms and acronyms or lexical hybrids that contain both letters and Chinese characters, and finally, numbers used as words. These types are all present in DCN. Overall, the neologisms in the dictionary can be classified into the following categories. Firstly, a large proportion of headwords that the dictionary records are influenced by other languages, particularly the English language. Most of these words are loan translations from English, as is shown in Table 1:

Table 1: Loan translations from English

Chinese headword	Source word in English	Chinese headword	Source word in English
暗网	dark web	第一先生	first gentleman
白金卡	platinum card	电子烟	e-cigarette
爆米花电影	popcorn movie	二手烟	second-hand smoke
边缘计算	edge computing	翻转课堂	flipped classroom
财政悬崖	fiscal cliff	间隔年	gap year
触屏	touch screen	口红效应	lipstick effect
词云	word cloud	路怒	road rage
大数据	big data	人盾	human shield

The second type of loanword is transliterations from English. Examples of this type include 艾特 (the sign @), 布基尼 (burkini), 玛丽苏 (Mary Sue), 慕客 (mook),

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1904

跑酷 (parkour), 披萨 (pizza), etc. In some entries, the source languages from which the headwords are derived are indicated right after the definitions, and such etymological information can be found in entries such as 乐活 (from English, LOHAS), 脸基尼 (from English, facekini), 尼特族 (from English, NEET), 罐头笑声 (from English, canned laughter), 轰趴 (from English, home party) and 提拉米苏 (from English, tiramisu). Neighboring languages such as Japanese and Korean have also contributed new lexical items to the Chinese language, as can be attested in headwords such as 达人 (expert, master), 卡哇伊 (kawaii), 森女 (mori girl), 手办 (garage kit), 宅男 (homebody) and 正太 (cute prepubescent young boy) from Japanese, 吃播 (mukbang), 韩流 (Hallyu, Korean wave) and 辛奇 (kimchi) from Korean. Another type of loanword that deserves mention is the so-called "letter words" (字母词), namely directly borrowed English initialisms and acronyms. The number of such borrowings has been on a steady increase, as can be seen from such entries in the appendix for words beginning with a Latin alphabet in the different editions of the authoritative *The Contemporary Chinese Dictionary*. The 1996 edition only recorded 39 such words while the number increased to 142 in the revision of 2002. In the sixth edition (2012), these words have increased to 241. They have been attracting more public attention due to the controversy over their exclusion in the A-to-Z part of Chinese dictionaries. DCN followed in the footsteps of the major general Chinese dictionaries, and put the list of neologisms beginning with signs, numerals and letters in an appendix which contains fifty pages and records more than 500 words. The majority of such words are English initialisms (e.g. ADSL, COO, O2O, UPS), and there are also combinations of numerals and Chinese characters, most of which should not be regarded as loanwords (e.g. "10后 person born after 2010," "B站 the website of Bilibili," "C字头 high-speed intra-city train beginning with the letter C").

Secondly, DCN includes over 50 Chinese characters, many of which are existing ones that have acquired new meanings. 二 (the numeral two), for instance, can now be used as an adjective meaning "stupid." Other frequently used old characters with new meanings are shown in Table 2. Some of them are rather productive as they can form many compounds.

Table 2: Characters with new meanings

Character	Old meaning(s)	New meaning
吧	used to soften the tone	bar, recreational facility
菜	vegetables, dishes	incompetent, inferior
处	place, department	director, department head
导	to lead, to guide	to direct, director
嗨	hi, hey	high, feeling euphoric

壕	moat, trench	uncultured nouveau riche
麦	wheat, barley	mike, microphone
萌	to bud, to sprout	cute, adorable
喷	to spurt, to gush	to criticize
扫	to clean with a broom, to sweep	to scan

刷 is a case in point. Originally meaning "to brush, to scrub," it now has obtained several new meanings like "to swipe (a card)," "to browse," "to be ID'd through scanning," "to write, to finish," etc. As a result, DCN recorded 27 compounds formed with 刷, such as 刷爆 (to exhaust the credit limit of one's bank cards), 刷脸 (to scan one's face), 刷屏 (to refresh the webpage or screen) and 刷题 (to do a large number of exam questions). There are also a few new Chinese characters coined mostly online in the dictionary. 囧, for instance, was previously used as an emoticon expressing embarrassment, helplessness, awkwardness, or surprise, and now it has not only become an established Chinese character, but also spawned several new compounds like 囧剧 (light comedy with embarrassing scenes), 囧片 (movie with awkward or embarrassing scenes), 囧事 (embarrassing matter), 囧态 (embarrassment), etc.

Thirdly, DCN has recorded hundreds of abbreviations or clippings. Unlike English, there are no initialisms or acronyms in Chinese. But to shorten a longer word has already become an ever-expanding trend, which may to some extent resemble clipping or blending in English. According to Zhang (2008: 267), shortened words account for 10.9% of all the new words she surveyed. In the past two decades, this kind of new words has greatly increased, which might be attributed to the factor that people tend to use shorter words not only on social media, but also on other informal occasions. Newly clipped Chinese words usually involve the deletion of one or more Chinese characters, as is indicated in the examples in Table 3.

Table 3: New clippings in Chinese

Clipped words	Full forms	English equivalent
半马	半程马拉松	half-marathon
超跑	超级跑车	supercar
创投	创业投资	venture capital
春招	春季招生	spring enrollment
单反	单镜头反光式取景照相机	single-lens reflex camera

电竞	电子竞技	eSports
电商	电子商务	e-commerce
黑五	黑色星期五	Black Friday
人盾	人肉盾牌	human shield
人肉	人肉搜索	to dox

Fourthly, characters or words that function like affixes or combining forms have played a key role in forming new derivatives or compounds. 族, originally meaning "tribe, clan" and "ethnic group," is used as a suffix-like character that can be attached to nouns, verbs, etc., referring to a group of people with common features. DCN did a fairly thorough job in including over 20 neologisms ending with 族, such as 乐淘族 (people enjoying shopping around), 留守族 (people who prefer to stay behind), 慢活族 (people living a simple and slow-paced life), 漂泊族 (people living a wandering life), 穷忙族 (the working poor), 穷游族 (people traveling on a budget), 银发族 (silver-haired people) and 月光族 (people who live paycheck to paycheck). Other characters or words that have formed at least four new words are indicated in Table 4.

Table 4: Productive characters or words

Character or word	Meaning	Examples
超级	super-	超级联系人 (supercontact), 超级细菌 (superbug), 超级月亮 (supermoon)
电子	electronic, e-	电子发票 (e-invoice), 电子护照 (e-passport), 电子垃圾 (electronic waste)
客	guest	拼客 (group buyer), 晒客 (person who shares things with others online), 职客 (person who charges a fee through providing job information)
秒	second	秒办 (to get it done quickly), 秒回 (to return a message immediately), 秒杀 (to seckill)
奴	slave	房奴 (person who works hard to pay off their mortgage), 孩奴 (parent who tries hard to satisfy their children's needs), 卡奴 (person who is unable to pay off their credit card debt)

拼	to share	拼车 (to carpool), 拼友 (person who shares the expenses)，拼租 (to rent together)
闪	flash	闪辞 (to resign shortly after being recruited), 闪赔 (to settle a claim immediately), 闪送 (to deliver within a short period)
微	micro-, Weibo or WeChat	微评 (to make comments on Weibo), 微商 (a WeChat business), 微塑料 (microplastics)
友	friend	车友 (fellow driver), 摄友 (fellow photograph buff), 微友 (fellow Weibo or WeChat user)
云	cloud	云媒体 (cloud media), 云课堂 (cloud classroom), 云平台 (cloud platform)

Fifthly, DCN also includes dozens of proper names ranging from company names to proprietary products. Those names that are well known internationally include 百度 (Baidu, a Chinese IT giant), 比特币 (Bitcoin), 谷歌 (Google), 脸书 (Facebook), 拼多多 (Pinduoduo, an online retailer), 欧冠 (European Championship), 淘宝 (Taobao, a Chinese e-commerce giant), 推特 (Twitter), 雅思 (IELTS) and so forth.

Finally, DCN records over 280 new meanings for existing characters, words and expressions. Although the Chinese language boasts many polysemous words, their number is definitely not comparable to that in the English language. However, in recent years we have seen a steady rise in new meanings created for existing words. A case in point is 八卦 (bagua), the eight divinatory trigrams according to *I Ching*, which can be used as an adjective, a verb and a noun, meaning "gossipy," "to gossip" and "gossip." Other examples are shown in Table 5.

Table 5: Words that have obtained new meanings

Headword	Original meaning	New meaning (s)
成绩单	report card, academic transcript	performance
干爹	adoptive father	sugar daddy
过山车	roller coaster	something characterized by unpredictable changes
恐龙	dinosaur	very ugly man
免费午餐	free lunch	something that is seemingly free of charge

皮肤	skin	customized graphic user interface
骑手	horseman, rider	motorcyclist, delivery man
钱包	wallet, purse	income, wealth

What makes DCN more readable and useful is its provision of rich etymological information and the three appendices. In some entries very detailed etymological information has been furnished, explaining in depth the origin of the headwords or providing background information, as is exemplified for neologisms such as 吃瓜群众 (onlookers, bystanders), 地球堂兄 (Earth's cousin), 共享经济 (sharing economy), 光盘行动 (clean your plate), 抗埃 (to fight against the Ebola virus) and 世遗 (World Heritage). The three appendices DCN has, containing more than 160 pages, are quite useful: the first one, as was previously mentioned, lists those frequently used neologisms beginning with signs, numerals and letters, which appeared after 1990s; the second one, lists over 600 hundred neologisms that were formed in the twentieth century and have not been recorded by *The Contemporary Chinese Dictionary*; the third one, a paper entitled "A survey on the use of neologisms in contemporary Chinese," sheds light on the use of 4 263 neologisms through analyzing their distribution, frequency, word-formational features, etc.

DCN is not immune to problems that often plague monolingual and bilingual neologisms dictionaries. The major deficiencies in DCN lie in the coverage of headwords and the provision of microstructural information. The problems with its headwords can be manifested in four aspects. Firstly, some of the headwords in DCN do not qualify as neologisms as they may be terms in certain fields or free combinations of at least two words, such as 大众旅游时代 (era of mass tourism), 独生子女护理假 (care leave for one's single child), 非首都功能 (functions nonessential to the role as the capital), 国际球员 (international player), 积木型住宅 (houses resembling building blocks), 浏览器主页劫持 (pagejacking), 马铃薯主粮化 (potato as a staple food), 农民安家贷 (loans granted to farmers to buy apartments), 舌尖上的腐败 (corruption through attending dinner banquets) and 隐形贫困人口 (invisible poverty-stricken population). Secondly, DCN has also committed the so-called "sin of omission," which is reflected in its failure to record two types of words: new words etymologically or semantically related to headwords already recorded and new words that are being frequently used. Examples of the first category include 钓鱼 (phishing, "钓鱼网站 phishing site" has been recorded), 扶贫 (to alleviate poverty, "扶贫云" has been recorded), 喷饭 (to split one's sides with laughter, "喷饭剧" has been recorded), 气候变化 (climate change, "气候债 climate debt" has been recorded), etc. As for the second category, notable absentees include 孵化器 (incubator), 加密货币 (cryptocurrency), 评论区 (comments section), 上线 (higher-level members of a pyramid scheme) and 网络诈骗 (Internet fraud, online fraud).

Microstructure-wise, DCN also leaves room for improvement, especially in its definitions, etymological information and labelling. Occasionally, the dictionary provides wrong definitions. 二面, for example, is interpreted as the second interview, but as the two illustrative examples show, it should refer to the second round of the interview. The definition for 碳补偿 (carbon offset) is also problematic as it has been considered a synonym for 碳中和 (carbon neutrality). According to ODE, carbon offset is defined as "an action intended to compensate for the emission of carbon dioxide into the atmosphere as a result of industrial or other human activity, especially when quantified and traded as part of a commercial scheme" while carbon-neutral is interpreted as "making or resulting in no net release of carbon dioxide into the atmosphere, especially as a result of carbon offsetting." Sometimes the wording for the definition may seem inappropriate, as is exemplified by what is provided for 键盘侠 (keyboard warrior) — a person who is cowardly and selfish in real life but acts righteously and makes bold comments online, which differs greatly from how ODE defines keyboard warrior, "a person who makes abusive or aggressive posts on the internet, typically one who conceals their true identity." The definition of 以太币 (Ethereum) also needs improvement. DCN defines it as a cryptocurrency that can be used to purchase virtual products and can be traded as well, but this definition can be applied to other cryptocurrencies like 比特币 (Bitcoin) and 泰达币 (Tether).

Etymologically, some of the headwords in DCN have been folk-etymologized. 奇异果, a transliteration of kiwifruit is said to be created because of the similar shape of Chinese gooseberry and the kiwi, a national bird of New Zealand. This may be pure conjecture as most English dictionaries including the OED do not provide any etymology for kiwifruit. Although *Wiktionary* provided detailed etymological information for the word, its explanation that the fruit got its name because its fuzzy brown skin resembles the plumage of the bird differs from what DCN describes. In the same vein, the etymology of 洗绿 (to greenwash) leaves room for improvement as it was not created in the same way as 洗钱 (money-laundering). To be precise, greenwashing, as is indicated in *Merriam-Webster's Collegiate Dictionary*, is a blend of "green" and "brainwashing."

The labelling in DCN also leaves a lot to be desired. The editor explains in the front matter that the dictionary adopts different policies of providing POS labels for words with three or more characters (e.g., three-character words being labelled on a case-by-case basis), which is unfriendly to users. The provision of labels for three-character words may seem haphazard. 霸王餐 is a case in point. It is interpreted as "do not pay one's meal at a restaurant, or can be likened to jawboning," but it is unlabeled, even though the phrase in which it is used (namely 吃霸王餐 "to dine and dash") is indicated after the definition. Moreover, DCN may have provided wrong POS labels for some headwords. 终面, for example, is defined as "the final round of interview," but is labeled as a verb. It is also quite strange to see that both 线上 (online) and 线下 (offline) are labelled as nouns. The only problem that can be identified with DCN's provision of illustrative examples lies in the fact that it provides examples showing wrong uses of the headwords. 霸凌 (to bully), for instance, is labelled as a verb, but its two illustrative examples showed its nominal use.

Dictionaries have always been viewed as the mirror of society, and dictionaries of neologisms may offer a better reflection of societal changes as they record the latest lexical changes. DCN, with its useful selection of Chinese neologisms and an abundance of illustrative examples, will undoubtedly serve its purpose of informing users who are interested in the lexical changes taken place in the Chinese language in the first two decades of the 21st century.

References

Dictionary Department, Institute of Linguistics, Chinese Academy of Social Sciences (Eds.). 2012. *The Contemporary Chinese Dictionary*. 6th edition. Beijing: The Commercial Press. (CCD)

Hou, M. 2016. *A Dictionary of Chinese Neologisms (2015)*. Beijing: The Commercial Press. (DCN)

Hou, M. 2023. *A Dictionary of Chinese Neologisms (2000–2020)*. Beijing: The Commercial Press.

Merriam-Webster Collegiate Dictionary. Springfield: Merriam-Webster. http://www.merriam-webster.com.

Oxford Dictionary of English. Oxford: OUP. http://premium.oxforddictionaries.com (ODE)

Oxford English Dictionary. 3rd edition. Oxford: OUP. http://www.oed.com (OED)

Yang, X. and W. Yang. 2009. An Analysis of Modern Chinese Neologisms. *Chinese Language Learning* 1: 97-104.

Zhang, X. 2008. *A Study on the Lexical Changes of Contemporary Chinese*. Jinan, Shandong: Qilu Publishing House.

Zhou, J. 2007. *A Dictionary of Chinese Neologisms (2006)*. Beijing: The Commercial Press.

Zou, Y. 2018. *A Dictionary of Chinese Neologisms (2017)*. Beijing: The Commercial Press.

Yongwei Gao
College of Foreign Languages and Literature
Fudan University
China
(ywgao@fudan.edu.cn)
(https://orcid.org/0009-0008-1532-8742)

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1906 (Project)

The Current State of the OBI DICT Project: A Bilingual e-Dictionary of Oracle-Bone Inscriptions with AI Image Recognition

Yang Jin, *Faculty of Humanities, Georg-August-Universität Göttingen, Germany (yangjin306@hotmail.com) (https://orcid.org/0009-0000-7353-9849)*
and
Shuo Wen, *MLBIO lab, Ecole Polytechnique Federale de Lausanne, Switzerland* (*shuo.wen@epfl.ch*) *(https://orcid.org/0009-0006-7923-7925)*

Abstract: This article reports on the current state of the OBI DICT project, a bilingual e-dictionary of oracle-bone inscriptions (OBI), incorporating artificial intelligence (AI) image recognition technology. It first provides a brief overview of the development of the lexicographical works on oracle bones. Subsequently, it identifies deficiencies in existing oracle-bone dictionaries and underscores the pressing demand for the compilation of a new dictionary. In the subsequent two sections, the article delineates the project's initiation and objectives and then outlines its design. The four principal phases of the project, that is, material collection, literature review, content and user interface design, and search engine and AI image recognition design, are described in detail in the third section. In the concluding section, it expounds on how the OBI DICT addresses users' search requirements and maximizes usability, thereby offering substantial support to contemporary oracle-bone research, streamlining the learning process for novices, and expanding the readership interested in oracle bones.

Keywords: ORACLE-BONE INSCRIPTIONS, ORACLE-BONE LEXICOGRAPHICAL WORKS, ORACLE-BONE DATABASES, BILINGUAL DICTIONARY, AI IMAGE RECOGNITION, MACHINE LEARNING, DICTIONARY COMPILATION

Opsomming: Die huidige stand van die OBI DICT-projek: 'n Tweetalige e-woordeboek van orakelbeeninskripsies wat gebruik maak van KI-beeldherkenning. In hierdie artikel word verslag gelewer oor die huidige stand van die OBI DICT-projek, 'n tweetalige e-woordeboek van orakelbeeninskripsies (OBI), wat kunsmatige intelligensie- (KI-) beeldherkenningstegnologie gebruik. Daar word eers 'n oorsig gegee van die ontwikkeling van die leksikografiese werke oor orakelbeendere. Daarna word leemtes in bestaande orakelbeenwoordeboeke geïdentifiseer en die dringende behoefte aan die samestelling van 'n nuwe woordeboek word beklemtoon. In die daaropvolgende twee afdelings word die ontstaan en doelwitte van die projek uiteengesit en daarna word die ontwerp daarvan beskryf. Die vier hooffases van die projek, nl. materiaalversameling, literatuurbeskouing, inhouds- en koppelvlakontwerp, en soekenjin- en

Lexikos 34 (AFRILEX-reeks/series 34: 2024): 198-212

KI-beeldherkenningsontwerp, word in die derde afdeling in besonderhede beskryf. In die slotafdeling word breedvoerig uiteengesit hoe die die soekvereistes van gebruikers in die OBI DICT aangespreek en die bruikbaarheid gemaksimaliseer word, om sodoende aansienlike steun aan kontemporêre orakelbeennavorsing te verleen, wat die leerproses vir leke vergemaklik, en die leserstal wat belangstel in orakelbeendere, uitbrei.

Sleutelwoorde: ORAKELBEENINSKRIPSIES, ORAKELBEEN- LEKSIKOGRAFIESE WERKE, ORAKELBEENDATABASISSE, TWEETALIGE WOORDEBOEK, KI-BEELDHERKENNING, MASJIENLEER, WOORDEBOEKSAMESTELLING

1. Introduction

The earliest unambiguously attested Chinese writing is the oracle-bone inscriptions (OBI), which refer to the texts inscribed on bones and shell in the Late Shāng (ca. 1300–1046 BC) and Western Zhōu (ca. 1046–770 BC) periods,[1] generally known as *jiǎgǔwén* 甲骨文 in Chinese. These texts were first excavated in 1899 in Anyang, China, and primarily constituted royal divinations. About twenty years after the discovery, the first lexicographical work on oracle bones appeared.[2] In the following one hundred years, oracle-bone materials were continuously unearthed, and more and more scholars joined the research, helping to achieve many significant accomplishments.[3] In this process, lexicographical works have also been continuously adapted to keep up with the development of the oracle-bone studies, and at the same time, in turn, have promoted the development of the discipline.

Since the discovery of oracle bones, there have been four kinds of oracle-bone (lexicographical) works, namely (1) collections of graphic forms, (2) concordances of inscriptions, (3) collections of lexical research, and (4) dictionaries. Collections of graphic forms focus on gathering various graphic forms of oracle-bone signs. Concordances compile inscriptions that feature the same oracle-bone sign. Collections of lexical research gather the perspectives of various scholars on the interpretation of oracle-bone signs. Dictionaries provide comprehensive analyses of the graphic forms and explain the usages of oracle-bone signs in the inscriptions. From the late 1980s, digital technology began to be integrated into oracle-bone research. The first online OBI database, the CHANT database (漢達文庫), was constructed in 1988 by the D.C. Lau Research Centre for Chinese Ancient Texts at the Institute of Chinese Studies, Chinese University of Hong Kong. Currently, the largest online OBI database is the Yīnqì wényuān 殷契文淵 (2016–), developed by Anyang Normal University, China. This database contains 154 collections of oracle-bone rubbings, including 239,736 images and 34,591 works on oracle-bone studies.[4]

With the assistance of digital technologies and online databases, the vast majority of research endeavours and projects on OBI have made significant progress. For instance, in the latest version of the *Xīn jiǎgǔwén biān* 新甲骨文編 (*New Collection of Oracle-bone Graphic Forms*) (Liú et al. 2014), digital technology

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1906 (Project)

is employed to realistically present oracle-bone glyphic forms, which can facilitate researchers' understanding of the oracle-bone graphic forms and their development in different periods. Online databases can also assist researchers in various ways. For example, database searches can identify inscriptions containing a specific oracle-bone sign within a few seconds, which enables them to serve as an electronic concordance. By combining searches in various online databases, researchers can efficiently compile a comprehensive concordance for a specific oracle-bone sign.

However, the development of oracle-bone dictionaries is apparently lagging behind. Although lexicographical works appeared not long after the discovery of oracle bones, the first oracle-bone dictionary, the *Jiǎgǔwén zìdiǎn* 甲骨文字典 (*Dictionary of Oracle-Bone Inscriptions*) (Xú 1988), was compiled in the late 1970s and was not published until 1988. Due to the influence of Chinese epigraphy, like most lexicographical works of ancient Chinese, this dictionary is handwritten and in vertical format. A revised version of this dictionary was published in 2022, with no significant alterations to its content but marked improvements in formatting. That is, the revised version no longer has a vertical layout in handwriting, but a horizontal layout in typeface. The oracle-bone dictionaries published after Xú (1988) primarily include: the *Xīnbiān jiǎgǔwén zìdiǎn* 新編甲骨文字典 (*Newly Compiled Dictionary of Oracle-Bone Inscriptions*) (Liú 1993, 2005), the *Jiǎnmíng jiǎgǔwén cídiǎn* 簡明甲骨文詞典 (*Concise Dictionary of Oracle-Bone Inscriptions*) (Cuī 2001), the *Yīnxū jiǎgǔwén shíyòng zìdiǎn* 殷墟甲骨文實用字典 (*Practical Dictionary of Oracle-Bone Inscriptions of Yīn Ruins*) (Mǎ 2008), and the *Shíyòng jiǎgǔwén zìdiǎn* 實用甲骨文字典 (*Practical Dictionary of Oracle-Bone Inscriptions*) (Chén 2019).

Among these dictionaries, some have better typesetting than Xú (1988), such as Mǎ (2008) and Chén (2019), which employ horizontal printing typesetting instead of vertical handwritten typesetting. This format presents no reading challenge for lay users as horizontal formatting aligns more with contemporary reading habits, which greatly reduce user difficulties and broaden readership. On the other hand, some of the layouts are not as good as Xú's (1988). For example, in Liú (1993, 2005), all kinds of information are listed together without obvious distinctions in each entry, which increases the difficulty of reading. Moreover, most of these latter dictionaries are not at the same level as Xú (1988) in explaining the usage of oracle-bone signs in the inscriptions. Let's take *mù* 目 as an example. The graphic form of this sign is the depiction of an eye. It has four usages in oracle-bone inscriptions, which are all listed in Xú (1988: 361-362, 2022: 248-249): (1) eye (n.), (2) spy, monitor (v.), (3) the name of a person, and (4) the name of a state,[5] while the same entry in Mǎ (2008: 88) and Chén (2019: 212) include only two usages. Besides, one of the usages in Mǎ (2008: 88), explaining the *mù* 目 as a name of sacrificial ceremony in oracle-bone inscriptions, is incorrect. Therefore, even though it has been published for more than thirty years, Xú (1988) is still the most widely used and authoritative dictionary in oracle-bone research.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1906 (Project)

However, the usages of oracle-bone signs in Xú (1988) reflect research from the late 1980s, failing to incorporate the research advancements made afterwards in the latest edition (Xú 2022). This inadequacy cannot satisfy the requirements of contemporary research. Therefore, the compilation of a better oracle-bone dictionary becomes an imperative necessity. Nowadays, the emergence of online oracle-bone databases and the application of interdisciplinary research approaches, such as the artificial intelligence (AI) image recognition technology combined with oracle-bone research method to recognize the oracle-bone signs (e.g. Huang et al. 2019, Liú 2020, Mén and Zhāng 2021, Jin 2023) or rejoin oracle-bone fragment images (e.g. Zhang et al. 2022), have brought new opportunities for dictionary compilation and oracle-bone research. If these new technologies and interdisciplinary research methods can be effectively utilized in the compilation of oracle-bone dictionaries, the newly compiled dictionary will undoubtedly promote the development of oracle-bone research and expand the readership of oracle-bone inscriptions.

2. Initiation and Aim of OBI DICT

The OBI DICT originated as part of a doctoral project at the Georg-August-Universität Göttingen, Germany (cf. Jin 2024). During the initial research phase of the project, it became evident that the development of oracle-bone dictionaries seriously lags behind oracle-bone research. Furthermore, it was noted that no English dictionary on oracle bones exists and other English reference works do not necessarily provide enough information on oracle-bone inscriptions. For example, the only comprehensive OBI introduction in English is Keightley's monograph published in 1978, which is now seen as outdated. English OBI readings published in recent years, such as Chen et al. (2017) and Takashima (2019), require high level of Chinese proficiency, which is not suitable or accessible for lay users. Based on the lack of user-friendly OBI dictionaries, we felt that there was a need for an accessible, modern bilingual oracle-bone dictionary. After investigating all previously mentioned oracle-bone lexicographical works and evaluating the feasibility of employing AI image recognition to the dictionary, the project started at Georg-August-Universität Göttingen at the end of 2022 (see also Jin and Wen 2024). Dr Yang Jin is the designer and coordinator of this project and Prof. Gordon Whittaker serves as the language consultant and academic advisor. Shuo Wen from the Ecole Polytechnique Federale de Lausanne provides AI technology support for the project. The project is expected to be completed within five years.

The digitalization of oracle bones, such as the construction of online databases, has rendered it more accessible to researchers and readers globally. Nonetheless, the effective utilization of these digital resources necessitates a high level of proficiency in the Chinese language, thus presenting a great challenge for both the general public with an interest in oracle bones and students in the early stages of their study worldwide. In addition to making up for the

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1906 (Project)

dictionary's shortcomings by adding the latest research results, the OBI DICT also intends to improve the dictionary's usability and expand its readership. In light of these challenges and opportunities, this project aims to develop a bilingual oracle-bone dictionary in the form of an electronic application using an interdisciplinary approach, including AI image recognition.

3. Design of the OBI DICT

The OBI DICT project consists of four main phases: material collection, literature review, content and user interface design, and search engine and AI image recognition design. The project has gone through the initial stage, that is, material collection and content design, and is now focusing on literature review and AI image recognition design.

3.1 Collection

The project has started with a comprehensive collection of academic works on oracle-bone inscriptions since late 1980s, including Chinese and Western sources. With regard to Chinese works, the collection is primary built on the collections of oracle-bone lexical research as well as online databases and platforms. Collections of lexical research primary include the (1) *Jiǎgǔ wénzì gǔlín* 甲骨文字詁林 (*Collection of Explanations on Oracle-Bone Inscriptions*) (Yú 1996), (2) *Bǎinián jiǎgǔxué lùnzhùmù* 百年甲骨學論著目 (*Bibliography of Oracle-Bones over the Past Century*) (Sòng and Cháng 1999), (3) *Jiǎgǔ wénxiàn jíchéng* 甲骨文献集成 (*Collection of Oracle-Bone Literature*) (Sòng and Duàn 2001), and (4) *Jiǎgǔ wénzì gǔlín bǔbiān* 甲骨文字詁林補編 (*Supplement of the Collection of Explanations on Oracle-Bone Inscriptions*) (Hé 2017). Yú (1996) is a collection of lexical research containing diverse scholarly perspectives pertaining to the interpretation of oracle-bone signs, extending from the discovery of oracle bones down to 1989. Sòng and Cháng (1999) is a catalog, which includes research works in China and abroad from the discovery of oracle bones in 1899 to June 1999. Sòng and Duàn (2001) is a collection of research works in China and overseas, covering the period from 1899 to 1999. Hé (2017) serves as a supplement to Yú (1996), which gathers the perspectives of various scholars from 1990 to 2013 regarding the interpretation of oracle-bone signs. Moreover, the latest research works primarily come from the continuously updated online oracle-bone databases, such as the Yīnqì wényuān 殷契文淵, and online platforms, such as the CNKI (中國知網). For example, the database Yīnqì wényuān 殷契文淵 contains 34,591 oracle-bone research works, which can be searched by the works' titles, authors, keywords, abstracts, sources, and full texts.

As for overseas research works, in addition to Sòng and Cháng (1999) as well as Sòng and Duàn (2001) mentioned above, overseas research after the late 1980s primarily come from: (1) the *Xīguānhànjì: Xīfāng hànxué chūtǔ wénxiàn yánjiū*

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1906 (Project)

gàiyào 西觀漢記: 西方漢學出土文獻研究概要 (*Chinese Annals in the Western Observatory: An Outline of Western Sinology's Contributions to the Study of Chinese Unearthed Texts*) (Shaughnessy 2018) and (2) the annual bibliography of the journal *Early China*. Shaughnessy (2018) is an historical introduction of Western research on Chinese unearthed materials, such as oracle-bone, bronze inscriptions, and bamboo and silk manuscripts. In the section of oracle bones, it has a list of Western research works from 1911 to 2015 (Shaughnessy 2018: 135-198). *Early China* is an annual journal, dedicated to original research covering every facet of China's culture and civilization, from ancient times to the Han dynasty (ca. AD 220). At the end of each volume, there is an annual bibliography, which is a list of English research works on early China for the whole year. Other research works on oracle-bones have also been gathered from online platforms such as Google scholar, ResearchGate and Academia.

3.2 Review

The collected works involve the interpretation of approx. 4,500 oracle-bone signs. Firstly, based on the (1) oracle-bone concordances, such as the *Yīnxū jiǎgǔ kècí lèi zuǎn* 殷墟甲骨刻辭類纂 (*Classified Compilation of Oracle-Bone Inscription from Yīn Ruins*) (Yáo 1989) and the *Yīnxū jiǎgǔ wéncí lèibiān* 殷墟甲骨文辭類編 (*Classification and Compilation of Oracle-Bone Inscriptions in Yīn Ruins*) (Chén 2021), and (2) online databases, such as the CHANT and Yīnqì wényuān 殷契文淵, the inscriptions for each individual oracle-bone sign are summarized. The collected research works have been studied, and the usage of each sign discussed in these works is summarized. In the next step, in order to determine the readable oracle-bone signs (approx. 2200) and their usage, the inscriptions are carefully analyzed, and the analysis is combined with the collected research works and the explanations provided in the dictionaries, such as Xú (1988, 2022), Cuī (2001), Liú (2005), Mǎ (2008), and Chén (2019). Then, appropriate inscriptional examples are selected for each usage. Finally, representative graphic forms for each sign are chosen based on previous collections of graphic forms, such as Liú et al. (2014) and Chén (2021), and the online database Yīnqì wényuān 殷契文淵.

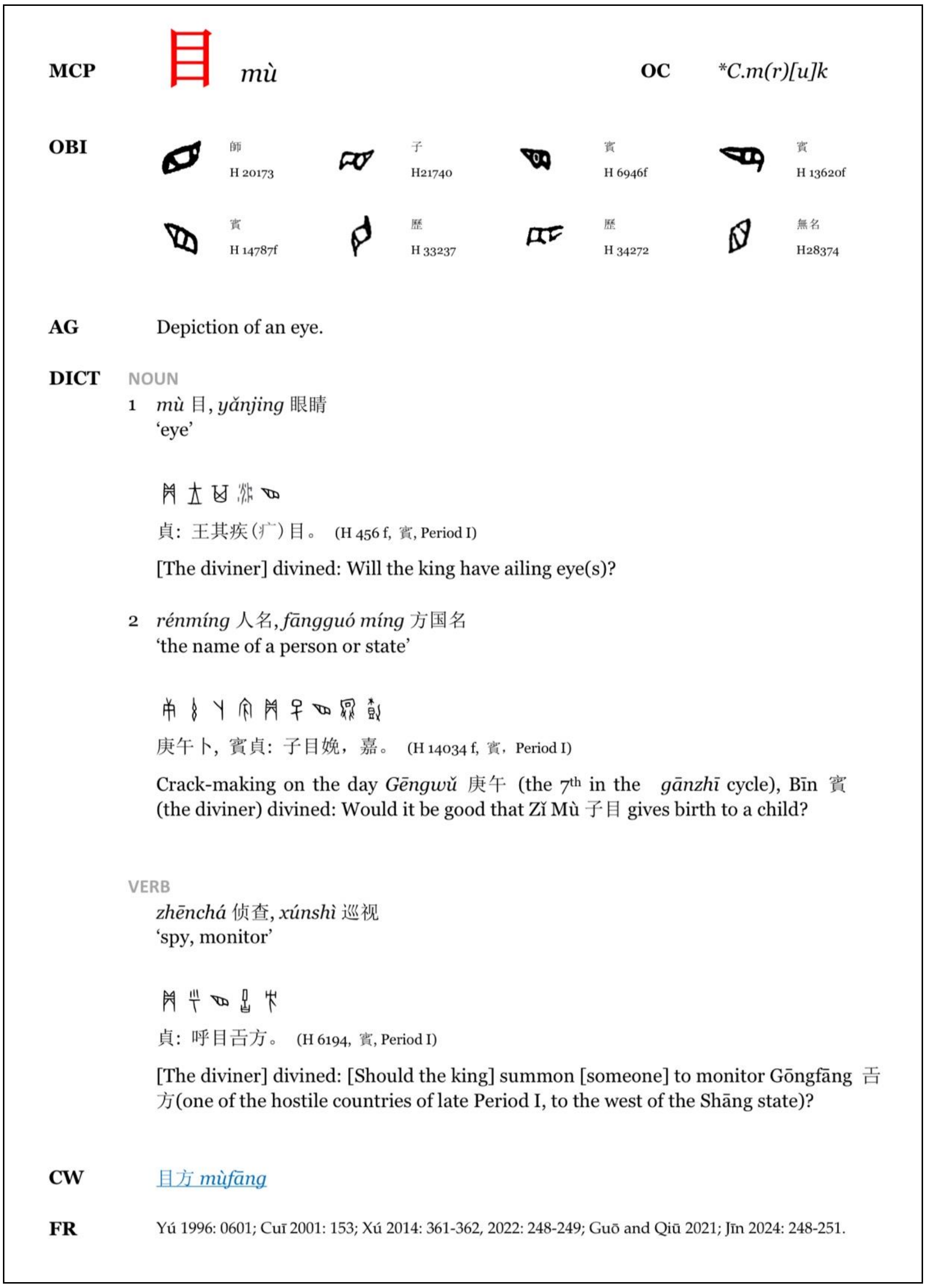
3.3 Design

3.3.1 Content

The OBI DICT contains approx. 2200 entries. Each entry includes, as illustrated by the entry *mù* 目 (see figure 1), the following elements:

(1) A head sign in Modern Chinese character with Pinyin (MCP):[6] e.g. 目 *mù*

(2) Old Chinese reconstructions (OC), where possible and available: e.g. **C.m(r)[u]k*



MCP 目 *mù* **OC** **C.m(r)[u]k*

OBI

師 H 20173
子 H21740
賓 H 6946f
賓 H 13620f
賓 H 14787f
歷 H 33237
歷 H 34272
無名 H28374

AG Depiction of an eye.

DICT NOUN

1 *mù* 目, *yǎnjing* 眼睛
‘eye’

貞：王其疾(疒)目。 (H 456 f, 賓, Period I)

[The diviner] divined: Will the king have ailing eye(s)?

2 *rénmíng* 人名, *fāngguó míng* 方国名
‘the name of a person or state’

庚午卜，賓貞：子目娩，嘉。 (H 14034 f, 賓，Period I)

Crack-making on the day *Gēngwǔ* 庚午 (the 7th in the *gānzhī* cycle), Bīn 賓 (the diviner) divined: Would it be good that Zǐ Mù 子目 gives birth to a child?

VERB

zhēnchá 侦查, *xúnshì* 巡视
‘spy, monitor’

貞：呼目舌方。 (H 6194, 賓, Period I)

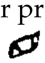
[The diviner] divined: [Should the king] summon [someone] to monitor Gōngfāng 舌方(one of the hostile countries of late Period I, to the west of the Shāng state)?

CW 目方 *mùfāng*

FR Yú 1996: 0601; Cuī 2001: 153; Xú 2014: 361-362, 2022: 248-249; Guō and Qiū 2021; Jīn 2024: 248-251.

Figure 1: The entry *mù* 目 in the OBI DICT

Pinyin is the modern pronunciation of this Chinese character, and the Old Chinese reconstruction is the pronunciation of the Shāng and Western Zhōu periods, which primarily follows the reconstructions of the Gassmann and Behr (2011), Baxter and Sagart (2014, 2020) and Zhèng (2018).

(3) Oracle-bone signs displayed in variant graphic forms with citations for their provenance from oracle-bone collections and diviner groups (OBI):[7]
e.g.  (H 20173, 師 group), (H 14787f, 賓 group)

One of the distinctive features of oracle-bone inscriptions, owing to its early developmental stage, lies in the variability of its graphic forms. In other words, a single oracle-bone sign may be represented in divergent graphic forms within the same historical period or in different periods. Therefore, it is imperative to compile a comprehensive list of these varied graphic forms, which aids readers in acquiring a thorough comprehension of the sign's graphic forms.

(4) An analysis of graphic forms (AG), as needed:

e.g. The graphic form of *mù* 目 is a depiction of an eye.

Oracle-bone inscriptions are comprised of numerous logograms, which serve as representations of lexical morphemes without explicit indication of word pronunciation. These logograms fall into two main categories: (1) those depicting objects or object parts, and (2) those depicting attributes, states, or actions.[8] Given this logographic nature, the analysis of graphic forms should be an essential part of oracle-bone dictionaries.

(5) The usage of the oracle-bone sign in inscriptions, along with English translations (DICT):

e.g. *mù* 目 is used as both a noun and a verb in inscriptions:
1. (n.) (1) *mù* 目, *yǎnjing* 眼睛 'eye'; (2) rénmíng 人名; fāngguó míng 方国名 'the name of a person or state'.
2. (v.) *zhēnchá* 侦查, *xúnshì* 巡视 'spy, monitor'.

(6) Illustrative examples of usage with provenance from oracle-bone collections and diviner groups. These examples consist of oracle-bone transcriptions, transliterations in traditional characters, and English translations:

e.g. One of the inscriptional examples for the usage (n.) (1) is:
貞王其疒目
貞: 王其疾(疒)目。 (H 456 f, 賓, Period I)
[The diviner] divined: Will the king have ailing eye(s)?

The inclusion of inscriptional examples is imperative for an oracle-bone dictionary, and providing citations that indicate the provenance of the examples is also of great importance. Such indications enable readers to revisit these inscriptions and form their independent assessments regarding the sign usage.

(7) References to related compound words (CW), including links to their usage:

e.g. The compound word related to *mù* 目 is *mùfāng* 目方.

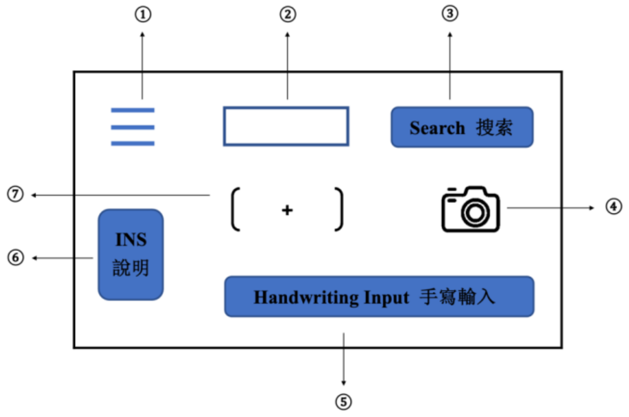
(8) Additional readings related to the specific oracle-bone sign (FR):

e.g. For more discussions on *mù* 目 see also Yú (1996: 0601), Cuī (2001: 153), Xú (1988: 361-362, 2022: 248-249), and Guō and Qiū (2021), Jin (2024: 248-251).

Similar to the indication of the examples' provenance, the additional readings related to the specific oracle-bone sign help readers establish a comprehensive understanding of the various discussions of the usage of the specific sign and form their independent views.

3.3.2 User interface

The oracle-bone signs are searchable by three modes in the OBI DICT: (1) Modern Chinese (both traditional and simplified), Pinyin or English through the search engine, (2) image (allowing users to upload pictures or take photos), and (3) handwriting input. These modes are designed to meet user search needs and improve usability to the largest extent (see figure 2).



① Settings
② Input box
③ Search button
④ Take a picture
⑤ Handwriting input
⑥ Instructions
⑦ Upload a picture

Figure 2: The user interface of the OBI DICT

Though professional oracle-bone researchers and users proficient in Chinese may comfortably employ Chinese characters or Pinyin for searching, it is also essential to meet needs of lay users and individuals less proficient in Chinese. The incorporation of AI image recognition (uploading pictures, taking photos and handwriting input) will significantly ease the retrieval and reduce the challenges associated with dictionary utilization for those not proficient in Chinese, greatly broadening the user base.

3.3.3 Maintenance and updates

One of the primary issues inherent in extant oracle-bone dictionaries, whether in print or digital form, is their failure to integrate the recent research findings. Therefore, the OBI DICT not only assimilates contemporary research results during its compilation but also commits to ongoing maintenance and updates post-completion, thereby aligning itself with the latest advancements in research. This practice guarantees the dictionary's status as a dynamic and up-to-date resource for users.

3.4 Methodology

There are mainly two parts in the method applied: (1) the image recognition model and (2) an OBI search engine. After the handwriting or photograph of an oracle-bone sign is input into the OBI DICT, it will be first recognized as a specific OBI sign by the image recognition model, and then the search engine will search and output the explanation of the sign (see figure 3).

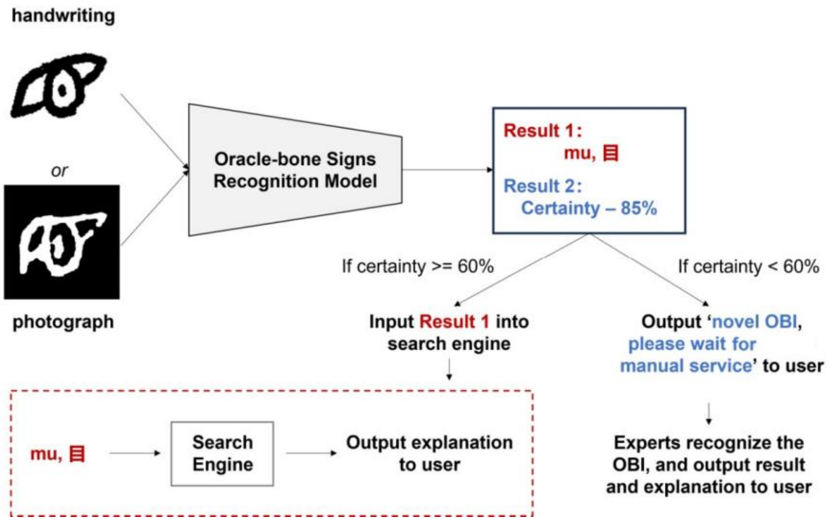

handwriting
or
photograph
Oracle-bone Signs Recognition Model
Result 1:
mu, 目
Result 2:
Certainty – 85%
If certainty >= 60%
If certainty < 60%
Input Result 1 into search engine
Output 'novel OBI, please wait for manual service' to user
mu, 目
Search Engine
Output explanation to user
Experts recognize the OBI, and output result and explanation to user

Figure 3: Recognition model and search engine of the OBI DICT

3.4.1 Image recognition

Machine learning methods are employed to identify oracle-bone signs in images. Specifically, identifying oracle-bone signs is considered as a classification task. In the first place, 2,200 predefined oracle-bone signs are used as 2,200 classes. Then, if input images contain oracle-bone signs, the model will output predefined classes representing the corresponding oracle-bone signs. Supervised learning is utilized to develop the classification model.

To train the model with supervised learning, a dataset must be created. Since the aim is to employ supervised training, both images and labels are required. To enable the model to recognize both photograph and handwriting of the OBI signs, images of both formats are collected. In short, the dataset consists of printed and handwritten OBI images accompanied by their respective labels. Let's again take the entry *mù* 目 as an example. We first collect 100 images of both photograph and handwritten format and resize them into 256 x 256 resolution. Then, 70, 20, and 10 images are put into training, validation, and test sets. The training sets are used for model training; the validation sets aid in hyperparameter selection, and the test sets assess the model's performance.

Our primary challenge lies in the scarcity of data available for robust model training. To address this issue, transfer learning should be employed. Broadly, transfer learning entails initializing the network using a large dataset that lacks oracle-bone signs, before fine-tuning the network using our OBI dataset. Another challenge is that the user can input rare OBIs which do not belong to the 2,200 predefined oracle-bone signs. In this case, out-of-distribution detection will be needed. Specifically, the model outputs a certainty for each input image. If the certainty is lower than a threshold, the model will output 'unknown OBI'. In this situation, if the user is willing to get further information, he or she can send a request and wait for assistance from a designated expert (manual service). The manual service can inform the user of the correct result or the information that this sign has not been deciphered yet.

3.4.2 Search engine

After identifying oracle-bone signs in images, results and explanations should be provided to the users. Thus, a search engine is needed to search the results and explanations according to the oracle-bone signs. The method for implementing the search engine is as follows: The dataset for the OBI DICT entries is in the first place constructed, including approx. 2200 recognized oracle-bone signs. Following that, a structured database is established using MySQL or PostgreSQL to store these entries. The database is then populated with entries from the DICT dataset, ensuring that each entry is inclusive of Chinese (both simplified and traditional), Pinyin, and English as searchable fields. Subsequently, the search logic on the backend is implemented, so that when a Modern Chinese character, Pinyin, or English input is provided by a user, the database is queried, and the corresponding entry is retrieved by the system. For example, to search for the oracle-bone sign corresponding to the word "eye", the user can enter the Modern Chinese character "目", Pinyin "mu" or English word "eye" in the search engine, and the OBI DICT will query the database and retrieve the corresponding entry *mù* 目 (see figure 1).

4. Conclusion

As discussed in section 1, the primary issue inherent in extant oracle-bone dictionaries is their failure to incorporate the latest research findings. Apart from that, there are some other shortcomings. For example, traditional handwritten and vertical formats present a tough reading challenge for lay users or non-native Chinese users. Moreover, the usages included in each entry are not comprehensive, which fails to enable users to understand the meaning of the same sign in different inscriptions. Furthermore, some dictionaries, such as Xú (1988, 2022) and Liú (1993, 2005), only have the lookup tables of the radicals and stroke counts,[9] which are usable for experts and users proficient in Chinese but are not friendly to lay persons and individuals less proficient in Chinese. Likewise, in most dictionaries, the pronunciation is not indicated, making it inconvenient to use for non-native Chinese users. The ongoing project OBI DICT has the potential to address all these problems. In the first place, it integrates the recent research findings and provides comprehensive usages for each entry and will also conduct an ongoing maintenance and updates to align itself with the latest advancements in research and provide up-to-date resource for users. Moreover, the OBI DICT employs horizontal typesetting instead of the traditional vertical handwritten format, which, as noted, aligns more with contemporary reading habits. Furthermore, the oracle-bone signs are searchable by three modes in the OBI DICT: Modern Chinese (both traditional and simplified), Pinyin or English through the search engine, image (allowing users to upload pictures or take photos) and handwriting input. The incorporation of AI image recognition or handwriting input will significantly assist the retrieval of data and reduce the challenges associated with dictionary utilization for those not proficient in Chinese. In addition, pronunciation for each entry, both Pinyin and Old Chinese reconstruction, Pinyin and English for the explanations of usages and English translations for illustrative examples are provided in the OBI DICT, making it user-friendly to a wider range of users. Therefore, the ongoing project OBI DICT aims to fully meet users' search needs and improve user usability to the largest extent. It provides support to current oracle-bone research, facilitates the learning process for beginners, and broadens the readership for oracle bones, including those whose first language is not Chinese.

The next stage of the OBI DICT project will focus on issues related to AI image recognition, such as improving the accuracy of image recognition, the recognition of undeciphered oracle-bone signs, and data privacy. In the maintenance phase, as mentioned in Section 3, the latest research on the approx. 2200 recognized oracle-bone signs will be continuously updated, and important discussions on approx. 2300 unrecognized oracle-bone signs will also be gradually added to the OBI DICT to facilitate users' research and learning. It would also be ideal if the OBI DICT could be a formal part of existing oracle-bone databases or platforms in the future. If the OBI DICT can be combined with these databases, it will greatly reduce learning difficulty, improve research efficiency, and promote the development of the discipline.

Endnotes

1. The chronology of the Late Shāng and Western Zhōu periods follows the latest research results of the Xia-Shang-Zhou Chronology Project (夏商周斷代工程) in China (Xià Shāng Zhōu duàndài gōngchéng zhuānjiāzǔ 2022).
2. The earliest lexicographical work on oracle bones is the *Fǔshì Yīnqì lèizuǎn* 簠室殷契類纂 (*Collections of Graphic Forms of Yīn Inscriptions in Fǔshì*), edited by Xiāng Wáng 王襄 and published in 1920.
3. For the development of academic research of oracle-bone inscriptions see e.g. Wáng and Yáng (1999), Wáng and Koo (2019) and Jin (2024: 68-74).
4. The database is still being continuously updated, and these are the statistics as of May 19, 2024.
5. For recent discussions on the oracle-bone sign *mù* 目 see Jin (2024: 248-251).
6. "Signs" refer to graphic forms representing Chinese before the Han Dynasty, when their graphic forms were not fully standardized, while "characters" refer to graphic forms representing Chinese after the Han Dynasty, when their graphic forms were standardized.
7. "H" is the abbreviation for the *Jiǎgǔwén héjí* 甲骨文合集 (*Collection of Oracle-Bone Inscriptions*) (Guō and Hú 1978–1982), the largest collection of OBI rubbings.
8. For recent discussions on the logograms in early Chinese writing see Jin (2024: 233-255).
9. Radicals, known as *bùshǒu* 部首 in Chinese, refer a component or a character conveying the lexical meaning of a logogram in Chinese, and Chinese dictionaries arrange characters under radicals.

References

Dictionaries

Chén, N. [陳年福] **(Ed.).** 2019. *Shíyòng jiǎgǔwén zìdiǎn* [實用甲骨文字典] (*Practical Dictionary of Oracle-Bone Inscriptions*). Chengdu: Sìchuān císhū chūbǎnshè [四川辭書出版社].

Cuī, H. [崔恒昇] **(Ed.).** 2001. *Jiǎnmíng jiǎgǔwén cídiǎn* [簡明甲骨文詞典] (*Concise Dictionary of Oracle-Bone Inscriptions*). Hefei: Ānhuī jiàoyù chūbǎnshè [安徽教育出版社].

Liú, X. [劉興隆]. 1993. *Xīnbiān jiǎgǔwén zìdiǎn* [新編甲骨文字典] (*Newly Compiled Dictionary of Oracle-Bone Inscriptions*). Beijing: Guójì wénhuà chūbǎn gōngsī [國際文化出版公司].

Liú, X. [劉興隆]. 2005. *Xīnbiān jiǎgǔwén zìdiǎn* [新編甲骨文字典] (*Newly Compiled Dictionary of Oracle-Bone Inscriptions*). Revised Edition. Beijing: Guójì wénhuà chūbǎn gōngsī [國際文化出版公司].

Mǎ, R. [馬如森]. 2008. *Yīnxū jiǎgǔwén shíyòng zìdiǎn* [殷墟甲骨文實用字典] (*Practical Dictionary of Oracle-Bone Inscriptions of Yīn Ruins*). Shanghai: Shànghǎi dàxué chūbǎnshè [上海大學出版社].

Xú, Z. [徐中舒] **(Ed.).** 1988. *Jiǎgǔwén zìdiǎn* [甲骨文字典] (*Dictionary of Oracle-Bone Inscriptions*). Chengdu: Sìchuān císhū chūbǎnshè [四川辭書出版社].

Xú, Z. [徐中舒] **(Ed.).** 2022. *Jiǎgǔwén zìdiǎn* [甲骨文字典] (*Dictionary of Oracle-Bone Inscriptions*). Horizontal Layout Edition. Chengdu: Sìchuān císhū chūbǎnshè [四川辭書出版社].

Other lexicographical works

Chén, N. [陳年福] **(Ed.).** 2021. *Yīnxū jiǎgǔ wéncí lèibiān* [殷墟甲骨文辭類編] (*Classification and Compilation of Oracle-Bone Inscriptions of Yīn Ruins*). Chengdu: Sìchuān císhū chūbǎnshè [四川辭書出版社].

Hé, J. [何景成] **(Ed.).** 2017. *Jiǎgǔ wénzì gǔlín bǔbiān* [甲骨文字詁林補編] (*Supplement of the Collection of Explanations on Oracle-Bone Inscriptions). Beijing: Zhōnghuá shūjú* [中華書局].

Liú, Z. [劉釗]**, Y. Hóng** [洪颺]**, X. Zhāng** [張新俊] **and Z. Zhōu** [周忠兵] **(Eds.).** 2014. *Xīn jiǎgǔwén biān* [新甲骨文編] (*New Collection of Oracle-bone Graphic Forms*). Revised Edition. Fuzhou: Fújiàn rénmín chūbǎnshè [福建人民出版社].

Wáng, X. [王襄]. 1920. *Fǔshì Yīnqì lèizuǎn* [簠室殷契類纂] (*Collections of Graphic Forms of Yīn Inscriptions in Fǔshì*). Tianjin: Tiānjīn bówùguǎn [天津博物館].

Yáo, X. [姚孝遂] **(Ed.).** 1989. *Yīnxū jiǎgǔ kècí lèi zuǎn* [殷墟甲骨刻辭類纂] (*Classified Compilation of Oracle-Bone Inscription from Yīn Ruins*). Beijing: Zhōnghuá shūjú [中華書局].

Yú, X. [于省吾] **(Ed.).** 1996. *Jiǎgǔ wénzì gǔlín* [甲骨文字詁林] (*Collection of Explanations on Oracle-Bone. Inscriptions*). Beijing: Zhōnghuá shūjú [中華書局].

Online databases

The CHANT Database (漢達文庫): http://www.chant.org/
The Yīnqì wényuān 殷契文淵 database: http://jgw.aynu.edu.cn/

Other literature

Baxter, W.H. and L. Sagart. 2014. *Old Chinese: A New Reconstruction*. New York: Oxford University Press. Online supplementary materials updated on October 21, 2020: http://ocbaxtersagart.lsait.lsa.umich.edu/

Chen, K., Z. Song, Y. Liu and M. Anderson (Eds.). 2017. *Reading of Shāng Inscriptions* [商代甲骨中英讀本]. Shanghai: Shanghai People's Publishing House.

Gassmann, R.H. and W. Behr. 2011. *Antikchinesisch — Ein Lehrbuch in zwei Teilen*. Bern: Peter Lang AG.

Guō, J. [郭靜云] **and S. Qiū** [邱詩螢]**.** 2021. Jiǎgǔwén zhōng yǐ tāotiè yǎnjīng wéi "mù" "chén" de zìxíng kǎo [甲骨文中以饕餮眼睛為"目""臣"的字形考] (A Study on the Graphic Forms of the Sign "Eyes" and "Officials" Based on Gluttons' Eyes in Oracle-Bone Inscriptions). *Jiǎgǔwén yǔ yīnshāngshǐ* [甲骨文與殷商史] 11: 280-291.

Guō, M. [郭沫若] **and H. Hú** [胡厚宣] **(Eds.).** 1978–1982. *Jiǎgǔwén héjí* [甲骨文合集] (*Collection of Oracle-Bone Inscriptions*). Beijing: Zhōnghuá shūjú [中華書局].

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1906 (Project)

Huang, S., H. Wang, Y. Liu, X. Shi and L. Jin. 2019. OBC306: A Large-scale Oracle Bone Character Recognition Dataset. *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, 20–25 September 2019:* 681-688.

Jin, Y. 2023. The Potential Benefits and Limitations of Artificial Intelligence Technology Used in Oracle-Bone Studies. *Irish Journal of Technology Enhanced Learning* 7(2): 8-20.

Jin, Y. 2024. *A Comparative Study of the Origins of Chinese and Mesoamerican Writing*. Ph.D. Dissertation. Göttingen: Georg-August-Universität Göttingen.
http://dx.doi.org/10.53846/goediss-10371

Jin, Y. and S. Wen. 2024. A Critical Analysis of OBI Lexicography and Online Databases, and a Brief Introduction to an Ongoing OBI e-Dictionary Project. (to appear)

Keightley, D.N. 1978. *Sources of Shang History: The Oracle-Bone Inscriptions of Bronze Age China.* Berkeley: University of California Press.

Liú, G. [刘国英]. 2020. Jīyú shēndù xuéxí de jiǎgǔ wénzì jiǎncè yǔ shíbié [基于深度学习的甲骨文字检测与识别] (Oracle-Bone Inscriptions Detection and Recognition Based on Deep Learning). *Yīndū xuékān* [殷都学刊] 3: 54-59.

Mén, Y. [门艺] **and C. Zhāng** [张重生]. 2021. Jīyú réngōng zhìnéng de jiǎgǔwén shíbié jìshù yǔ zìxíng shùjùkù gòujiàn [基于人工智能的甲骨文识别技术与字形数据库构建] (Artificial Intelligence Based Oracle-Bone Inscriptions Recognition Technology and Graphic Forms Database Construction). *Zhōngguó wénzì yánjiū* [中国文字研究] 1: 9-16.

Shaughnessy, E.L. 2018. *Xīguānhànjì: Xīfāng hànxué chūtǔ wénxiàn yánjiū gàiyào* [西觀漢記: 西方漢學出土文獻研究概要] (*Chinese Annals in the Western Observatory: An Outline of Western Sinology's Contributions to the Study of Chinese Unearthed Texts*). Shanghai: Shànghǎi gǔjí chūbǎnshè [上海古籍出版社].

Sòng, Z. [宋鎮豪] **and Y. Cháng** [常耀華] **(Eds.).** 1999. *Bǎinián jiǎgǔxué lùnzhùmù* [百年甲骨學論著目] (*Bibliography of Oracle-Bones over the Past Century*). Beijing: Yǔwén chūbǎnshè [語文出版社].

Sòng, Z. [宋鎮豪] **and Z. Duàn** [段志洪] **(Eds.).** 2001. *Jiǎgǔ wénxiàn jíchéng* [甲骨文献集成] (*Collection of Oracle-Bone Literature*). Chengdu: Sìchuān dàxué chūbǎnshè [四川大學出版社].

Takashima, K. 2019. *A Little Primer of Chinese Oracle-Bone Inscriptions with Some Exercises.* Second revised edition.Wiesbaden: Harrassowitz.

Wáng, Y. [王宇信] **and S. Yáng** [楊昇南]**.** 1999. *Jiǎgǔxué yǐbǎinián* [甲骨學一百年] (*One Hundred Years of Oracle-Bone Studies*). Beijing: Shèhuì kēxué wénxiàn chūbǎnshè [社會科學文獻出版社].

Wáng, Y. [王宇信] **and Y.H. Koo** [具隆會]. 2019. *Jiǎgǔxuéfāzhǎn 120 nián* [甲骨學發展120年]. (*120-Years' Development of Oracle-Bone Studies*). Beijing: Zhōngguó shèhuì kēxué chūbǎnshè [中國社會科學出版社].

Zhang, Z., A. Guo and B. Li. 2022. Internal Similarity Network for Rejoining Oracle Bone Fragment Images. *Symmetry* 14(7): 1464.

Zhèng, Z. [鄭張尚芳]. 2018. *Shànggǔ yīnxì* [上古音係] (*Ancient Phonological System*). Shanghai: Shànghǎi jiàoyù chūbǎnshè [上海教育出版社].

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1907

Heming Yong, Jing Peng and Xiangming Zhang. *Chinese Lexicography in the Twentieth Century.* 2024, x + 280 pages. ISBN 978-1-6366-7529-9 (Hardback), ISBN 978-1-6366-7527-5 (eBook PDF), ISBN 978-1-6366-7528-2 (ePub). New York/Berlin/Brussels/Chennai/Lausanne/Oxford: Peter Lang. Price: US$ 94.95 (Hardback), US$ 90.20 (eBook)

The history of lexicography published around the globe is generally approached from two perspectives. One is to explore the history from an international perspective, mostly concerning the development of lexicography in English-speaking countries (e.g. Béjoint 2010); while the other is to adopt a domestic stance, investigating the history of a specific country (e.g. Yong et al. 2009). A large number of these works, however, sketch the general development of lexicography, and pay little attention to detailed information over a certain period. The current volume *Chinese Lexicography in the Twentieth Century* stands out from others of the same kind. It sheds light on the evolution of Chinese lexicography within a 100-year span and elaborates upon the development of major dictionary types in China from both the sociocultural and linguistic points of view.

This volume consists of seven chapters. The first five chapters respectively survey Chinese philological dictionaries, Chinese bilingual dictionaries, Chinese special and encyclopaedic dictionaries, Chinese learner's dictionaries, and Chinese electronic dictionaries thoroughly. The sixth chapter depicts how lexicography made progress in Hong Kong, Macao, and Taiwan, compared with that in the mainland. The last chapter, as sublimation of those achievements illustrated in the previous chapters, gives an overview of the theoretical explorations in Chinese lexicography.

Chapter one documents the rapid development of Chinese philological dictionaries, featuring the transition from traditional patterns to modern ones with modern linguistic attributes. This chapter begins with the presentation of the sociocultural circumstances that exerted profound effects on Chinese lexicography, such as the trend of "Chinese learning as essence, western learning for practice," the May Fourth Movement, the Anti-Japanese War, the policy for "illiteracy elimination," the Great Cultural Revolution and the Reform and Opening-up Policy. It then explicates how Chinese philological dictionaries went through three stages, from the innovation and transformation (1900–1949) to the renascence and retrogression (1949–1976) and to the stabilisation and flourishing (1976–1999) by means of the exemplary works. *The Great Character Dictionary of the Chinese Language* (《中华大字典》1915) is recognised as epoch-making in modern Chinese lexicography, in terms of its innovative endeavour to systematically include Chinese characters, simplify the system of phonetic notation, differentiate senses and homographs, etc. This chapter pays special attention to Chinese dialect dictionaries as well. After introducing their origins and popular classification, it analyses how the compilation of such dictionaries evolved from individual-motivated behaviours to nation-supporting projects, from the empirical study to the data-based research, and from distributed models to systematic ones. This chapter concludes with a summary of the characteristics and prospects of Chinese philological dictionaries.

Chapter two focuses on the revolutionary advancement of Chinese bilingual dictionaries in the 20th century. It emphasises the correlation between social environments and the quantity and quality of bilingual dictionary compilation. Though experiencing several social upheavals which almost led to the halt of the compilation, Chinese bilingual dictionaries finally came to thrive at the end of the 20th century and play a dominant role in the market. In the first half of the century, the frequent military aggressions and unfair treatments in the international world stirred up public fervour to learn from the Western world and Japan. Consequently, a number of English–Chinese and Japanese–Chinese dictionaries were published at the same time. Moreover, a number of Russian–Chinese dictionaries were produced when China and Russia formed strong bonds after 1949. Subsequently, the Chinese bilingual dictionary explored its way from mere simulation to the absorption and digestion, and to the original compilation. *The English–Chinese Dictionary (unabridged)* (《英汉大词典》 1989–1991), with LU Gusun as the editor-in-chief, epitomised the maturity and unprecedented success of Chinese bilingual dictionary-making in that it adopted such creative techniques as the objective descriptive paradigm, the self-built database, its independent compilation, and the distinct macrostructure. Apart from bilingual dictionaries with Chinese and foreign languages, this chapter also expounds on how the scale of bilingual dictionaries with Chinese ethnic languages expanded with wider diversity and larger numbers.

Chapter three illustrates how special and encyclopaedic dictionaries as well as encyclopaedias gained momentum from 1900 to 1999. Before listing and analysing the major works, this chapter discusses the social and historical backgrounds at length, including the socioeconomic conditions for those types, the history of Chinese special and encyclopaedic dictionaries, the origin of "encyclopaedia," etc. It considers the sophistication of science and technology as the prerequisite for the advancement of special and encyclopaedic dictionaries. It then moves on to segment each type of dictionaries into different periods of time with reference to the domestic movements. In general, special, and encyclopaedic dictionaries as well as encyclopaedias share a similar track — the initial-development stage before the 1950s, the depression stage during the 1950s and the 1970s, and the accelerating stage after the 1970s. This chapter ends with a summary of the characteristics and prospects of the special and encyclopaedic lexicography.

Chapter four ushers in the concept of "learner's dictionaries" which was completely new for China before the twentieth century. As the learner's dictionary is a relatively young type in lexicography, this chapter investigates its historical background first. The theoretical findings from the 1920s to the 1930s, including those of descriptivism, phraseology, and grammar, laid a solid foundation for the emergence of learner's dictionaries whose targeted readers are English as second or foreign language learners. Under the three pioneers' leadership

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1907

(i.e. Harold Edward Palmer, Michael Philip West, and Albert Sydney Hornby), English learner's dictionaries manifest unique qualities. Learner's dictionaries are characterised with the timely application of the latest linguistic theories and technologies, the selective entry coverage, abundant grammatical instructions, multi-functional examples, and the control of defining vocabularies. Next, this chapter enumerates the footprints of English monolingual learner's dictionaries, Chinese–English learner's dictionaries, and Chinese monolingual dictionaries. Significant terms (e.g. the "Big Four") and groundbreaking dictionaries are elucidated, contributing to a broad overview of this topic. This chapter ends with the deficiencies and prospects of Chinese learner's dictionaries in the 20th century.

Chapter five examines the role computer technology played in Chinese dictionary-making in the twentieth century. Overall, computer science was burgeoning in the last two decades of the century, thus giving birth to electronic dictionaries. Resembling the development of Chinese bilingual dictionaries, Chinese electronic dictionaries followed the order from inception, simulation, integration, to rapid development. To make a comprehensive description, this chapter starts with a survey of the typology and characteristics of electronic dictionaries. Then it unfolds the advancement of Chinese electronic dictionaries in a whole-to-part manner. Notably, it was Taiwan and Hong Kong that advanced and assumed the key roles in the expansion of Chinese electronic dictionaries since the first Chinese electronic dictionary was generated in Hong Kong and the first Chinese CD electronic dictionary was made in cooperation with Taiwan. The last section reminds readers about some deficiencies of Chinese electronic dictionaries and gives directions of the path ahead.

Chapter six reports on the 20th-century Chinese lexicography in Hong Kong, Macao, and Taiwan. As the places that were subject to invasions and occupation and once separated from mainland China, they all conformed to their own developmental paces and attributes. As for Hong Kong, limited varieties and number characterised the early years of the century. After the 1950s, the rapidly expanding population and accelerating economic growth gave rise to a gradual increase in dictionary production. After the 1970s, with the establishment of excellent publishing companies, Hong Kong became a global base for Chinese publication, promoting the making of large-scale dictionaries. In contrast to Hong Kong, Macao, which boasted a rather long and glorious history of communication with Western civilisation, suffered from a decline in terms of the dictionary publication. This can be ascribed to the 20th-century downtrend to the weakening status of Macao for international communication compared with the past. The 20th-century lexicography in Taiwan was concomitant with the political turmoil. The Japanese occupation from 1900 to 1945 saw the prevalence of Taiwan dialect-Japanese dictionaries; from 1945 to 1987, notwithstanding the strict ideological control of the Kuomintang government, Chinese lexicography still made its way forward with better quality and quantity; after 1987, driven by the lift of martial law and the communication with the mainland China, lexi-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1907

cography made greater strides in multi-aspects, such as a wider coverage of themes and more varieties published.

The last chapter of this volume provides an overview of lexicographical theories in China. It probes into the theoretical inquiries into Chinese lexicography during the 20th century. It describes in detail the ways of elevation from the Chinese dictionary-making to the academic research, the general trend towards systematic and in-depth explorations as well as corresponding events. Following this, it discusses the features in the foundation, construction and formation stages of inquiries, the debates over the status of lexicography as an independent discipline, studies in various branches of Chinese lexicography, and application of linguistic findings to lexicography. Furthermore, this chapter delineates a picture of lexicographic academic activities during the twentieth century, involving the establishment of academic teams and organisations, the contributions of publishing companies and journals, and the organisation of domestic associations and international conferences.

Overall, the volume under review is distinguished from other similar publications in at least three aspects. Firstly, different from other monographs on Chinese lexicography, which centre on the whole history spanning over hundreds of years from the origin to the modern time, this volume narrows its time span of explorations down to the twentieth century. As a result, in addition to the inquiries into those dictionaries with profound and archetypal impacts, others with pivotal values are listed as well, delivering a more panoramic view of the developmental path. Secondly, this volume is innovative in the incorporation of the sociolinguistic paradigm into the illustration of lexicographical history. No field of study could survive without the social environment, and lexicography is of no exception. This work demystifies the fluctuations of the 20th-century Chinese lexicography in a broader context of political, cultural, and economic changes, the inherent bonds with Chinese traditions, and the interdependence of the international backgrounds. Thirdly, the methods of the demonstration in this volume are distinct. Contrary to the regular plain statement, it repeatedly supports claims with convincing statistical evidence. Besides, the content of every influential dictionary is discussed exhaustively, covering the selection of entries, the retrieval system, sense differentiation, definitions, quotations, phonetic notations, as well as the external perspectives, like the motive of compilation, the prestige, the influence, and the process of revisions.

Aside from the aforementioned virtues, there is still some room for this volume to attain excellence. There are a few minor mistakes, such as the misspelling of some dictionary editors' names (e.g. "Liang Shih-chiu" is incorrectly spelt as "Liang Qiushi"), and inconsistencies in the symbols and names of Chinese dictionaries (e.g. there are two different English renderings of 《词诠》). In Chapter Five, the definition of electronic dictionaries is lengthy, and may easily perplex readers. Additionally, the structure would be more coherent if there were a concluding chapter which summarises the main findings of previous chapters and discusses the implications and avenues for further research.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1907

To conclude, this deserves recognition as the first volume on the history of the 20th-century Chinese lexicography applying both sociolinguistic and lexicographical theories. It depicts a dynamic, objective, comprehensive and detailed path of how Chinese lexicography, motivated by the torrents of the challenging times, has developed from its traditional mode to the modern, scientific, linguistics-related, and systematic one. This volume is a timely addition to the literature of history of lexicography.

Acknowledgements

This study was supported by the National Philosophy and Social Sciences Foundation of China (Grant No. 22BYY010).

References

A. Dictionaries

Hornby, A.S., E.V. Gatenby and H. Wakefield (Eds.). 1942. *Idiomatic and Syntactic English Dictionary.* Tokyo: Kaitakusha.

Lu, G. (Ed.). 1989–1991. *The English–Chinese Dictionary (unabridged).* Shanghai: Shanghai Translation Publishing House.

Palmer, H.E. (Ed.). 1938. *A Grammar of English Words*. London: Longmans, Green & Co.

West, M.P. and J.G. Endicott (Eds.). 1935. *New Method English Dictionary.* London: Longmans, Green & Co.

Xu, G., P. Ouyang and C. Wang (Eds.). 1915. *The Great Character Dictionary of the Chinese Language*. Shanghai: Zhong Hua Book Company.

B. Other Literature

Béjoint, H. 2010. *The Lexicography of English: From Origins to Present*. Oxford: Oxford University Press.

Yong, H., J. Peng, B. Tian and X. Zhang. 2009. *Chinese Dictionaries: Three Millennia (From 1046 BC to AD 1999)*. Shanghai: Shanghai Foreign Language Education Press.

Yong, H., J. Peng and X. Zhang. 2024. *Chinese Lexicography in the Twentieth Century.* New York: Peter Lang.

Yuanwen Zhang
Centre for Linguistics and Applied Linguistics
Guangdong University of Foreign Studies
Guangzhou
China
(20232510003@gdufs.edu.cn; yuanwen-zhang@qq.com)
(https://orcid.org/0009-0009-4956-7888)

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1908 (Article)

Grammatical Data in the *Dictionary of Montenegrin National and Literary Language*

Sonja Nenezić, *Faculty of Philology, University of Montenegro,*
Nikšić, Montenegro (sonjan@ucg.ac.me)
(https://orcid.org/0009-0005-7711-3139)

Abstract: Dictionaries are generally consulted to ascertain the meaning of a word. However, the meaning is inseparable from its grammatical features, which often determine it. Therefore, this article examines the type, scope, and method of presenting grammatical data in a comprehensive general dictionary. The aim is to analyze and ascertain the morphological and syntactic characteristics of all types of words recorded in what is currently the only such dictionary of the newly standardized Montenegrin language. Attention is also given to the applied metalanguage, representing a combination of transparent abbreviations and natural language. The initial hypothesis about the heavy reliance on the inherited Serbo-Croatian lexicographic practice is confirmed, but certain deviations from this tradition are also noted, which aligns with the dictionary's goal of presenting grammatical data more accessibly and comprehensively to its target users.

Keywords: GRAMMATICAL DATA, GRAMMATICAL MARKER, GENERAL DESCRIPTIVE DICTIONARY, *DICTIONARY OF MONTENEGRIN NATIONAL AND LITERARY LANGUAGE*, SERBO-CROATIAN LANGUAGE DICTIONARIES

Opsomming: Grammatikale data in die *Dictionary of Montenegrin National and Literary Language*. Woordeboeke word gewoonlik geraadpleeg om die betekenis van 'n woord te bepaal. Die betekenis kan egter nie van die grammatikale kenmerke, wat dikwels die betekenis daarvan bepaal, geskei word nie. Daarom word die tipe, bestek, en metode van aanbieding van grammatikale data in 'n omvattende algemene woordeboek in hierdie artikel bestudeer. Dit het die analisering en bepaling van die morfologiese en sintaktiese eienskappe van al die tipes woorde wat tans in die enigste sodanige woordeboek van die nuut gestandaardiseerde Montenegrynse taal opgeneem is, ten doel. Aandag word ook geskenk aan die toegepaste metataal, wat 'n kombinasie van deursigtige afkortings en natuurlike taal verteenwoordig. Die aanvanklike hipotese dat daar sterk gesteun word op die oorgeërfde Serwo-Kroatiese leksikografiese praktyke word bevestig, maar sekere afwykings van hierdie tradisie word ook waargeneem, wat ooreenstem met die doel van die woordeboek om grammatikale data meer toeganklik en omvattend vir sy doeltaalgebruikers aan te bied.

Sleutelwoorde: GRAMMATIKALE DATA, GRAMMATIKALE MERKER, ALGEMEEN BESKRYWENDE WOORDEBOEK, *DICTIONARY OF MONTENEGRIN NATIONAL AND LITERARY LANGUAGE*, SERWO-KROATIESE TAALWOORDEBOEKE

Lexikos 34 (AFRILEX-reeks/series 34: 2024): 218-235

1. Introduction

It is well acknowledged that in dictionary-making, lexicographers primarily focus on the meanings of lexical units, but the grammatical features inherent to them are equally crucial for their usage. Proponents of integral linguistic theories, with Apresjan (2010) being one of the most notable, advocate for the integration of dictionaries and grammar towards a unified linguistic description, emphasizing the need for their mutual consistency in terms of the data included and the method of its recording. Achieving this requires the collaboration of their authors, applying the same theoretical approaches and principles of "identification, classification, and interpretation of linguistic units," which is quite rare (Topolinjska 2002: 33). Typically, a dictionary and a grammar, as two fundamental works describing a language, where the former lists lexical items and the latter prescribes the rules for their combination, are produced independently of each other. Moreover, there is an issue where grammar and other linguistic disciplines do not provide suitable solutions for dictionary compilation, and conversely, the extent to which lexicography is receptive to the existing linguistic literature is discussed (Tafra 2005: 167). However, one thing is certain: linguists, even those not advocating for specific integral models, recognize the necessity and importance of grammatical information in the lexicographic processing of lexemes (Gortan-Premk 1980, Katičić 1994).

The central place in a dictionary is indeed reserved for lexical data, as it is most commonly consulted to verify the meaning of words (Engelberg and Lemnitzer in Kostić-Tomović 2017: 21). The question arises as to how much grammar should be included and by what principles, to ensure that the lexicographical description is as precise and purposeful as possible. It is essential to remember that although a dictionary is complementary to grammar, it is still a separate entity, distinct from grammar, hence it should contain a certain amount of grammatical data, without which it would otherwise be unusable or very difficult to use (Kačić 1994: 302). This necessary minimum must be primarily determined for any large monolingual descriptive dictionary, especially one intended for a diverse range of users such as native speakers, foreign users who have a (greater or lesser) command of the language, and experts — e.g. lexicographers using it as a basis for developing other types of dictionaries (Kostić-Golubičić 1997: 458). Therefore, our aim in this article is to analyze the interrelationship between grammar and the dictionary, exemplified by the only such lexicographic work of the newly standardized Montenegrin language to date. The lexicographic work in question is the *Dictionary of Montenegrin National and Literary Language* (DMNLL) published by the Montenegrin Academy of Sciences and Arts, in which the lexemes, as the subject of description, are presented "in the entirety of its grammatical forms and meanings" (DMNLL 2016: IX).

For more political than linguistic reasons (see Šubarić and Đurčević 2023), work on this fundamentally important lexicographical work was suspended immediately after the release of the first volume. The long-prepared second volume has therefore not yet been published. This, of course, causes delays in

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1908 (Article)

establishing lexical norms for Montenegrin (Šubarić and Đurčević 2023: 72), and there is uncertainty on when and how this issue will be resolved. However, since it is currently the only general dictionary of the Montenegrin language, it is still in use, and linguists, despite its incompleteness, utilize it in their research, especially since it will undoubtedly serve as a basis for many different academic lexicographical endeavors. We believe that it is necessary to promptly address the identified and other potential linguistic shortcomings, to which no lexicographical work is immune, finalise the second edition, and continue working on this crucial project for Montenegro, which is why we have carried out this research. The evaluation of the grammar presented in the DMNLL will take into account the type and amount of grammar as well as how the grammatical data is presented. The impact of grammar on the meaning of the headword will also be considered.

Montenegrin[1] was standardized after Serbian, Croatian, and Bosnian, following the dissolution of Serbo-Croatian as the common language of Serbs, Croats, Bosniaks, and Montenegrins, triggered by the disintegration of the Yugoslav state community. It is presumed that the authors of the DMNLL heavily relied on inherited Serbo-Croatian lexicographical practices. Therefore when indicating the lexicographic treatment of lexemes in the DMNLL, reference will occasionally (when it is necessary to highlight differences) be made to the lexicographical treatment of grammatical data in the most relevant multivolume dictionaries of Serbo-Croatian: The *Dictionary of Serbo-Croatian Literary and National Language*[2] (1959–) (DSCLNL) and the *Dictionary of Serbo-Croatian Literary Language*[3] (1967–1976) (DSCLL). However, considering the temporal distance between DSCLL and the early volumes of DSCLNL, on the one hand, and DMNLL, on the other hand, it is expected that the DMNLL is somewhat more contemporary and purposeful.

Some brief remarks on the DMNLL and a general discussion on grammatical data as a part of lexicographical data and its representation in a general monolingual dictionary follow. Subsequently, the morphological and syntactic features which are included in the lexicographical description of headwords in the DMNLL are investigated. Additionally, we will examine whether there are deviations in the presentation of grammar from the inherited Serbo-Croatian lexicographical practice, and what these deviations are. We will highlight any shortcomings and omissions that should be addressed in the continued work on the development of the Montenegrin dictionary. The most significant findings of our research will be succinctly presented in the conclusion.

2. *Dictionary of Montenegrin National and Literary Language* — general remarks

The basic information about the corpus of our research can be found in its preface (DMNLL: IX-XII). In 2011, the Montenegrin Academy of Sciences and Arts established the Council for the Compilation of the Montenegrin Language Dic-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1908 (Article)

tionary, which included experts from various fields. Aiming to encompass the diverse lexicon of the Montenegrin area, the Council opted for the title *Dictionary of Montenegrin National and Literary Language,* which is also rooted in lexicographical tradition (cf. *Dictionary of Serbo-Croatian Literary and National Language* of the Serbian Academy of Sciences and Arts). This dictionary represents a collective endeavor in all phases of its creation: from the selection of sources, the excerpting of lexical material, to its processing. The first volume of the DMNLL, printed on the tenth anniversary of the restoration of Montenegrin statehood, contains 12,018 words listed in alphabetical order, starting with the Cyrillic letters A, B, and V. It was initially planned for this dictionary to have around 100,000 entries. It is conceived as a general descriptive dictionary, hence the lexicon it covers includes, in addition to the commonly used lexicon of the literary language, dialectal, terminological, onomastic, and all lexicon that is temporally and expressively marked with any semantic or morphological peculiarity. Pertinent to our study is the emphasis that it reflects current literary language norms at all levels of language structure — orthographic, orthoepic, grammatical and lexical, but at the same time, it can serve as a supplement to it, contributing to the resolution of dilemmas present in prosody, orthography, phonology, morphology, word formation, and syntax. Arguably the dictionary is therefore not only descriptive, but also prescriptive[4].

3. Grammatical data in a general monolingual dictionary

The amount of grammatical data to be included in a dictionary and the lexicographic solutions to be applied depends largely on the type of dictionary and the target audience it is intended for, the characteristics of the language whose lexicon is being presented, and the lexicographic tradition (cf. Atkins and Rundell 2008: 399, Zgusta 1991: 115-116). In some dictionaries, typically bilingual ones, where the focus in processing the headwords is on equivalents, grammatical overviews of one or both languages are often provided among its supplementary sections, although the grammatical data do not always need to be so comprehensive and can be limited to specific key categories, such as tables of irregular verbs (Kostić-Tomović 2017: 45-46). However, in general monolingual dictionaries, which aim for comprehensiveness, much more data, not only about grammatical, but also about semantic, pragmatic, and other properties of the headword can be found within the dictionary entry itself.

Grammatical data is often presented or contained in the comment on form, comprising the morphological and syntactic features of the headword and having a normative character. Morphological features pertain to parts of speech, their grammatical categories, inflection, and word formation. Syntactic features, on the other hand, include syntagmatic relations, primarily valency, especially of verbs, but also of other morphological classes, the functions of individual words, etc. It should be noted that the system of parts of speech and their categories is equally important for syntax, hence it is also referred to as the morpho-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1908 (Article)

syntactic system. It is a fact that traditional Serbo-Croatian lexicography devotes more space to morphological than to syntactic data, as highlighted by one of the more significant Yugoslav and Serbian syntacticians, Popović (2003: 204), who emphasizes that it deals with words, not the syntactic units formed from them, and does not pay enough attention to the syntactic aspects of lexeme usage. However, it is clear that some types of words, such as conjunctions and prepositions, can only be defined by their function. The greater prevalence of morphological data in the dictionaries of many Slavic languages is influenced by their intricate inflection, within whose paradigms various accentual and morphological deviations from the canonical form of headwords occur. In English, for instance, inflection is less developed, and its paradigms are generally quite predictable, hence they are usually seldom displayed lexicographically. This is why, for example, Atkins and Rundell (2008: 218-221) do not mention it in *The Oxford Guide to Practical Lexicography*, where they highlight three types of grammatical data: (1) word class, (2) constructions or syntactic relationships, specifically of nouns, verbs, adjectives, and adverbs, as these are the four main word types, and possibly (3) data that directly depends on the headword's class, such as countability for nouns, indication of action or state for verbs, attributive or predicative function for adjectives, etc.

In addition to other elements, grammatical data contributes to the value of a dictionary and should meet some basic requirements: it should be unambiguous, complete, and economical (Gruszczyński in Kostić-Golubičić 1997: 458). It could be said that the grammatical description of the lexicon illustrates the relation between macrogrammar and microgrammar. This is especially reflected in the presentation of exceptions because grammar is, in principle, "more oriented towards rules, the system, analogy [...], while the dictionary records individuality, anomalies" (Tafra 2005: 68). A user will consider a dictionary the best if it includes everything they are looking for. Besides the basic forms, users are often, for example, interested in other inflected forms, grammatical peculiarities of a word, etc. Since every dictionary has its grammatical scheme, it should also contain, usually in the introductory section, some kind of guide through it, which would include grammatical markings and explain how they are used (Atkins and Rundell 2008: 218).

When discussing indicators of grammatical data, it is important to remember that establishing a metalanguage is one of the major lexicographic challenges. Lexicographers from different countries theoretically consider and practically resolve the issue of creating a universal metalanguage by using semantic primitives and the symbolic means of their denotation. They aim to construct interconnected interpretations of words from different languages. However, in domestic lexicography, preference is given to the natural language for accurate and consistent dictionary interpretations (Kozyrev and Černjak 2015: 74). Therefore, lexicographic data, including grammatical ones, can be encoded and decoded. Most printed dictionaries use abbreviations, while in electronic dictionaries, grammatical terms are usually given in full (Atkins and Rundell 2008: 218). Codes can be transparent, opaque with multiple keys, and opaque with a single

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1908 (Article)

key (Kostić-Tomović 2017: 47). Transparent ones are based on abbreviations and symbols that the target user group is already familiar with, having encountered them during their education, such as abbreviations for cases, parts of speech, etc. Codes whose meaning we cannot infer from prior knowledge or from the context are considered opaque. They are not desirable for the user even when used repeatedly with the same meaning, let alone singly, when they must be deciphered anew each time. Often, we learn about the grammatical properties of a headword indirectly, through illustrative examples. One example can convey more than ten symbols that the user does not understand (De Caluwe and Van Santen 2003: 82).

The grammatical markers themselves can be direct or indirect indicators of a grammatical category (Gortan-Premk 1980: 108-109). For instance, in traditional Slavic dictionaries, the direct marking of masculine, feminine, or neuter gender indirectly indicates the association of the headword with nouns. Although such an approach ensures the economy of grammatical data, there is an increasing demand for it to be explicit nowadays. This is the case, for example, in electronic English dictionaries (Atkins and Rundell 2008: 219), as well as in Apresjan's *Active Dictionary of the Russian Language* (2010). A dictionary is expected to play a more active role, and the lexicographical data provided should be as comprehensive and direct as possible, thereby becoming more accessible and useful to the user. Finally, it is worth noting that various typographical and nontypographical structural markers play a significant role in presenting grammatical data: the sequence of markers, different types of fonts, the use of various kinds of brackets, etc.

4. Grammatical data in the *Dictionary of Montenegrin National and Literary Language*

The creation of the DMNLL is based on the *Handbook for Processing Entries in the Dictionary of Montenegrin National and Literary Language*, which establishes a unique practice for the lexicographic processing of the selected lexemes. This handbook provides exemplary descriptions for all types of words with a complete physiognomy of the dictionary entry, which includes the way of presenting the headword, its pronunciation form, the sequence of grammatical data, etymology, syntagmatic and phraseological expressions, the method of determining meanings, the choice of appropriate examples, and the listing of sources. The mentioned *Handbook*, as well as the dictionary's macrostructure, which includes the *Preface, Tabular Overview of Lexicographical Processing* (XIII-XVII), *Instructions for Using the Dictionary* (XIX-XXVII), *Abbreviations* (XXIX-XXXII), and *Symbols and Punctuation Marks* (XXXIII), provide a good insight into the lexicographical solutions related to the scope and manner of presenting grammatical data in the DMNLL, although some of them were modified and refined during the creation process. The use of abbreviations and symbols, especially pronounced in the grammatical processing of headwords, indicates a consideration for the economy of lexicographical description, as it pertains to a print edition. It should be noted that these are transparent codes, mostly traditionally established, and there is also information that is not coded. The grammatical block in the dictionary entry comes after the possible pronunciation form, which is provided along with the standard accented form of the headword, and etymological data, but grammatical data, especially if related to particular meanings of headwords, also appears in the section reserved for the definition (see the examples in Appendix A).

It is entirely logical that from a grammatical perspective, different types of words are not processed in the same way; that is, the grammatical data is conditioned by the type of word. If it belongs to the variable word classes, the headword, as the first segment of the dictionary entry, is presented in its basic, canonical form, which represents the entire paradigm (Zgusta 1991: 115). For nominal words (nouns, adjectives, pronouns, and numerals) in Montenegrin, as in many other languages, this is the nominative singular, less often plural, and for verbs, the infinitive. Invariable words: prepositions, conjunctions, interjections, and particles, appear only in one form, and therefore they are accompanied by a marker of the morphological class to which they belong. The same lexicographic treatment is applied to adverbs, although some of them, mainly adverbs of manner, have comparative forms. This solution ensures their differentiation from neuter gender adjectives, with which many adverbs coincide in form (e.g., **brzo** (*fast*) as both adverb and neuter gender adjective), but also from other word types that derive from them (e.g., **više** (*more*) as both an adverb and a preposition). Among the variable word types, nouns, adjectives, numerals, and verbs are presented without a type marker, which is indicated through grammatical categories or definitions. Pronouns are an exception, presumably primarily because of the adjectival pronouns, which users commonly confuse with adjectives, and also because they are classified similarly in English grammar. Generally, the same practice is traditionally applied in Serbo-Croatian and in Slavic dictionaries at large, whereas in English dictionaries, each word type is assigned its qualifier since verbs and nouns often share the same form, and adjectives do not acknowledge gender categories, thus eliminating any ambiguities that might arise for the dictionary user (Marković 2014: 71-72).

In the subsequent sections, we will analyze the representation of grammatical features of all word types in the DMNLL, except for numerals which are not included in the first volume, whilst also highlighting certain deviations from the DSCLL and DSCLNL.

4.1 Nouns

According to traditional lexicographic practice, nouns do not carry a mark of the class they belong to. Instead, they are equipped with data on grammatical categories such as gender, number, and case, which are inherent to them (Čirgić et al. 2010: 68). However, it is important to note that the data about the word type is part of the typical grammatical definition for verbal nouns, for example,

abdiciranje (*abdicating*),[5] a verbal noun derived from the verb "abdicirati" (to abdicate).

Nouns have the most comprehensive grammatical data regarding the category of gender. Alongside all of them, the gender is initially indicated by an abbreviation: "m." (masculine), "ž." (feminine — henceforth f.), or "s." (neuter — henceforth n.). In the DSCLNL and DSCLL, however, the singular genitive form is provided before the gender, if necessary, without a case marker, while the other cases, introduced by an abbreviation, are listed after the gender mark and in round brackets. Since grammatical gender, unlike with adjectival words, is a classification category for nouns, its markers indirectly suggest that their bearers belong to the noun class of words. Therefore, this sequence in the DMNLL seems to be a better solution.

Two grammatical gender markers appear alongside pairs of headwords that differ in their endings, for example, **brigàntīn** and **brigantína** m. and f. (a light sailing ship with two masts), but each variant corresponds to one gender and paradigm. There are also rarer cases where one form has two genders without a difference in paradigm and meaning, for instance, **bijenále** (biennial) n. and m., and also with a noted difference, e.g., **bôl** m. and f. 1. (usually in masculine) *the feeling of physical pain due to injury or illness* 2. (usually in feminine) *the feeling of mental pain; sorrow, sadness*. However, the markers m. and f. refer not only to grammatical but also to natural gender or sex. This dictionary, unlike the DSCLNL and DSCLL, attempts to separate and mark the grammatical and natural gender of nouns where they do not coincide. Without delving into the complex issues of distinguishing gender and sex here, although it should be noted that here grammatical gender is treated as a morphological, not a syntactic category, we have observed several types of nouns that receive two gender markers, specifically when: (a) forms of the masculine gender denote people of both male and female sex (nomina professionalis), e.g., **akadèmik** (academician) m. (+f.), where the female form **akàdemkinja** is also recorded; (b) forms of the feminine gender denote people of female and male sex (nomina attributiva), e.g., **avétinja** (fool) f. (m.); (c) they refer to a person or animal of male sex, but are grammatically of feminine gender, e.g., **burègdžija** (a person who makes burgers) m. (gram. f.); (d) variants of masculine and feminine genders represent male and female persons, e.g., **ànglist(a)** (anglicist) m. (f.), with the female form **ànglistkinja** also presented. In the DSCLNL and DSCLL, the mentioned nouns only receive the first of the two indicated markers, although there are examples of common-gender nouns, mainly nomina attributiva, with two markers, e.g., **budala** (fool). However, inconsistencies are observed in the DMNLL as well, as, for instance, the headwords **analìtičar** (analyst), **ankètar** (polister), **apsòlvēnt** (graduate) etc., are marked only as masculine nouns even though they are used for female persons as well.

Regarding declension, the data about it is nowhere near as exhaustive as in the case of the gender category. Not all case forms of nouns are noted, primarily for efficiency reasons. Moreover, it is debatable whether it is necessary to occupy dictionary space with listings of patterns that users already know or

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1908 (Article)

can easily predict (De Caluwe and Van Santen 2003: 82). Therefore, nouns either lack any data on case forms, or, grouped in square brackets, only those suffixes or complete case forms of nouns that differ in accent or morphology from the basic form are noted. These forms deviate from systematic predictability, and the case abbreviation informs about which form of the word is in question, e.g., **bàstadūr** [*gen. sg.* bastadúra, *voc. sg.* bȁstadūre, *instr. sg.* bastadúrom] (a person who is resourceful, skilled; everything is going their way). Although such practice is also present in Serbo-Croatian dictionaries, it is noticeable from a general overview that the DMNLL justifiably devotes more space to case forms, especially in comparison with the DSCLL. For the mentioned lexeme, for example, both the DSCLL and DSCLNL provide only the genitive singular ending.

Duplex forms are also noted, separated by a slash, e.g., **alkohòličār** [*voc. sg.* -e/-u, *instr. sg.* -om/-em] (alcoholic). Rare nouns that lack declension, mostly loanwords, are marked with the qualifier "*nepromj.*" (invariable), e.g., **Bȁntu** (Bantu)**, vònderbra** (wonderbra) ... The case forms of nouns that change meaning when inflected thus gain the status of a separate defining word, for example, **večȉnōm** (mostly).

When it comes to grammatical number, it is well-known that common nouns, besides being listed in the singular, also have a plural form, which is usually not noted except in cases of accentual or morphological deviation from the basic form. In such instances, under the label "*nom. pl.*" (nominative plural), its full form or just the ending is specified, either alone or together with other case forms, e.g., **àviopark** [*nom. pl.* aviopàrkovi] (airpark), **bȁmbrēk** [*voc. sg.* bȁmbrēče, *nom. pl.* bȁmbrēci] ... (a short fat person; chunky person). In the DSCLNL and DSCLL, however, only the abbreviation "*pl.*" is used, without a case marker, even when the plural nominative form is provided along with other case forms.

If a noun has two accentual and/or morphological plural forms, both are listed and separated by a slash, for example: **vȅlegrad** [*nom. pl.* velegràdovi/ vȅlegrādi] (metropolis), **buldožèrist(a)** [*gen. sg.* -a/-ē, *nom. pl.* -i/-e] (catskinner). A noun that occurs only in the plural is marked with the Latin abbreviation "*pl. t.*" (pluralia tantum), e.g.: **Apeníni** pl. t. (The Apennines). We consider this designation to be more precise than the *mn.* label used in the analyzed Serbo-Croatian dictionaries. However, nouns predominantly used in the singular, which are mostly material and abstract, including verbal nouns, are not marked as such either in the DMNLL, DSCLNL or DSCLL, even though such data would be valuable for users.

Only collective nouns receive a special qualifier: "*zb.*" (italicised) if they are collective both morphologically and lexically (e.g., **bȁlavčād** *zb.* from *balavče*), and "zb." (not italicised) if they are collective only lexically (e.g., **balavùrdija**). This unusual difference in font style is not noted in the list of abbreviations. If a noun is not collective in all its meanings, then the "zb." label is placed before the specific sense, for example, **bižutèrija a.** zb. *jewelry made from cheap materials and fake stones*. This indirectly indicates that these lexemes or sub-lexemes do not have plural forms. On the other hand, as a suppletive form, for example, the plural of the noun **brȁt** [*supl. pl.* brȁća] (brother) is marked, which is also provided as a separate defining word with the "*zb*." qualifier, a more user-friendly approach as it does not require grammatical knowledge from the user.

The grammatical category of number can significantly impact the meaning of a lexeme, as seen in the case of zoological and botanical terms, where the plural form denotes a species, while the singular form refers to an individual representative. For example, **àlbatros** in plural refers to *oceanic birds from the Diomedeidae family with a large wingspan*; in singular, it *denotes such a bird.* **Bèscvjetnica** in singular means *a plant that lacks flowers or does not bloom*; in plural, *it refers to such plants, known as Cryptogamae*. Other instances of this type are very rare, but we can illustrate with the lexeme **vrijème** (time), one of whose meanings is related only to the singular: **1.** (in sg.) **a.** philos. *one of the two dimensions of existence (alongside space), manifested in the form of continuous duration.*

Many nouns, either entirely or in a specific meaning, are accompanied by data about their colligation tendency[6] towards the plural form, for example, **bjȅčva** (usually in pl.) refers to a *type of short sock or leg warmer*, **bjèlača** 1. (usually in pl.) *short socks made of white wool*, **bàtina** 2. (only in pl.) *received beatings: to get a beating.*

The DMNLL also notes specific functions of nouns, which are enumerated as separate senses, for example: **vozàrica 1.** *a female carter*, **2.** (in attributive use) *propelled by rowing, with oars* (about a boat): *boat ~*; **vrućȋna 1.** *high air temperature, scorching heat.* **2.** (in adverbial use) *hot, excessively warm*; **automàtik 1.** *a device or machine that independently performs the action for which it is designed*. **2.** (in the role of an invariable adjective) *operating automatically, independently: ~ gearbox, etc.* Data about the noun's rection (see **batàljōn**, battalion etc.) is not provided, neither in the DMNLL, nor in Serbo-Croatian dictionaries.

4.2 Adjectives

Like nouns, adjectives do not have a part-of-speech marker, except for invariable adjectives (marked with the qualifier "*nepromj. pridj.*"), which are almost all of foreign origin, like the adjective **bàtāl** (cannot be used anymore, outdated) from Turkish and **blȉnd** (which protects, shields) which is from German. The canonical form of variable adjectives is the nominative singular masculine form, in the indefinite aspect if it exists, and in the positive, if it features a category of degree.

Gender is fully represented, as with nouns, so the forms for feminine and neuter genders are always provided, specifically their endings if they are distinguished only by these, or their full forms or occasionally complete forms if they accentually and morphologically deviate from the base form, e.g., **vòlovskī**, -ā, -ō (bovine), **bèščūlan**, -lna, -lno (senseless), **bȉjel**, bijèla, bijèlo (white). This indirectly indicates the morphological class of the defining word. With rare adjectives, data is provided in round brackets indicating that they are predominantly or exclusively used in the feminine gender, either entirely or in one of its senses, e.g., **brȅmenit**, -a, -o **1.** (only in fem. gender) *meaning pregnant.*

Data on case and number is not provided because there are no deviations from the usual paradigm; instead, data on aspect and degree are provided in square brackets. Specifically, for all adjectives that have forms in both aspects, the definite aspect form is given after the marker "*odr.*", alongside the canonical indefinite form, for example, **ȁlāv**, -a, -o [*odr.* ȁlavī] (a person who eats a lot, insatiable, voracious, greedy), whereas in the DSCLNL and DSCLL, this is only done if they deviate in accent or phonemes from the base form. The dependency of a particular meaning of an adjective on its definite aspect is also regularly noted, e.g., **vìsok**, visòka, visòko **10. a.** (only in def. aspect) *which refers to someone in a prominent position in service; related to persons in such positions*, as well as the tendency of certain adjectives to favor this aspect, for instance: **vȉšeglasan**, -sna, -sno and **višèglasan**, -sna, -sno [*odr.* vȉšeglasni and višèglasni] (usually in def. aspect) *composed of multiple voices; performed with several voices, polyphonic*. Adjectives that have only the indefinite or only the definite aspect do not carry information indicating the absence of forms in both aspects, although the latter indirectly signify this fact through their canonical forms.

Comparative forms are also provided in square brackets, but only, as in the DSCLL, if there have been phonetic alterations in their formation, for example, **bȉjel**, bijèla, bijèlo [*comp.* bjȅlji] *which is the color of milk, snow*. In contrast, in the DSCLNL, they are always recorded. Suppletive comparatives have their own dictionary entry, for instance, **bȍljī**, -ā, -ē *comp. of good.* The conditionality of a certain sub-meaning by the tendency towards the comparative form is noted as well, e.g., **vìsok**, visòka, visòko **6. b.** (usually in comp.) *aspiring to something noble, sublime, significant; possessing such qualities: higher ideals; higher principles*. Superlative forms are justifiably not listed among the forms or as defining words, since their formation follows the rules without deviation.

Data about the rection with which certain adjectives are distinguished is noted by listing complements in the form of indefinite pronouns for people and things in the corresponding case, provided in round brackets before the definition of the lexeme or sub-lexeme, e.g., **vȉčan** (to something) **a.** *accustomed to something, used to something*. **b.** *skilled in something, experienced, adept, proficient.* **c.** *well-acquainted with something, informed about something*; **vjȅšt 3.** *(at something) having great knowledge about something, very adept at something, accustomed to something, skilled, etc*. There are also omissions in listing, as the data on rection is missing, for example, with the adjective **blagonáklon** (benevolent, favorable).

Noun usage of adjectives, which is often accompanied by restrictions related to aspect or number, is regularly noted as a separate meaning/sense. For example: **bȍgāv**, -a, -o **2.** (in nominal usage) m. *poor soul, weakling*; **vȍljen**, -a, -o **2.** (in nominal usage) (in definite aspect) *a person who is loved*; **blîžnjī**, -ā, -ē and blȉžnjī, -ā, -ē **3.b.** (in nominal usage) (usually in pl.) *a person who is in close kinship with someone, a relative, kin; a person in general.* However, certain substantivized adjectives have separate dictionary entries, e.g., **Bȕgarskā** (Bulgaria), indicating a lexicographic issue with the unclear boundary between polysemy and conversion.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1908 (Article)

4.3 Pronouns

Alongside pronouns, a qualifier for their morphological class ("*pron.*") is provided, and in the definition, sub-classes are specified, while characteristic case forms are given in square brackets: including accentual forms, enclitic forms, and even those characteristic of dialects, marked with the abbreviation "*nar.*". For example: **vî** pron. [*gen.* vâs, encl. vas, *dat.* vȁma, encl. vam, nar. vi, *acc.* vâs, encl. vas, nar. ve] **1.** *personal pronoun for the second person plural; the people to whom the speaker is addressing in speech or writing*. **2.** (Vi) *for polite addressing of an individual, as an expression of respect: thank you.* **3.** (in dat.) *functioning to enhance the meaning and draw attention*. Adjective pronouns, as they are characterized by motion, are listed in the masculine gender, with endings for feminine and neuter, and with data about their class and subclass, e.g., **vȁš**, -a, -e *possessive pronoun for 2nd person pl.* **1.** *belonging to the larger number of people we address (to you).* **2.** (Your) *in addressing someone out of respect, meaning "your".* **3.** (in nominal usage) m. (in pl.) *family members, relatives; like-minded individuals, members of a party or society, etc.* From these examples, we see that data is provided not only about the formative characteristics but also about the usage of the pronouns.

4.4 Verbs

Verbal entries are presented in the infinitive as the basic form, without a morphological class marker. Rare verbs that lack an infinitive are listed in the first person present tense and marked with the abbreviation "*incomp.*" (incomplete verb), for example, **vèlīm** (complete and incomplete) [pres. *vèlīš, vèlī,* imperf. v*ȅljāh*] incomp. meaning to *say, tell, speak*.

Information about the verbal aspect, which can be perfective (*svrš.*) (perf.) or imperfective (*nesvrš.*) (*imperf.*), is a mandatory grammatical detail. These markers also serve as indirect indicators of the word type. Verbs of different aspects but the same meaning are listed as separate entries, e.g., **búpati** (imperf.), **bȕpiti** (perf.) (to hit something making a dull sound, smash). Verbs with two aspects, predominantly of foreign origin, receive both markers: perf. and imperf., for instance, **blokírati** (block)**, vizuèlizovati** (visualize). Subsequently, in square brackets, paradigmatic forms are registered, typically the 1st person singular present tense, and if there is a duplication of forms, then the 3rd person plural, as well as other forms if they deviate accentually or morphologically from the headword. For instance, alongside the verb **vȉdjeti** (see), besides the present tense, the imperative, imperfect, active participle, passive participle, past adverbial participle, and present adverbial participle are provided, indicated by abbreviations: [*pres.* vȉdīm, *imp.* vȉdi, *imperf.* vȉđāh, *act. part.* vȉdio, vȉdjela, vȉdjelo, *pass. part.* vȉđen, -a, -o, *past adv. part.* vȉdjēvši, *pres. adv. part.* vȉdēći]. The DSCLL records only the suffix for the 1st person singular present tense, without a marker, while the DSCLNL additionally provides the aorist, but lacks the passive participle and adverbials. For impersonal verbs, marked with the abbreviation '*impers.*', the mandatory present tense in the grammatical block is provided in the 3rd person singular, which is the only form used, for example, **bjȅnūti** (perf.) [pres. bjȅnē] impers. *meaning to clear up without full brightness, to become slightly beautiful, to improve (regarding the weather).*

The dictionary entry for verbs that appear in both non-reflexive and reflexive forms, that is, without and with the particle "*se*", is divided into two parts marked with Roman numerals I and II. The first part deals with their non-reflexive use, and the second part with their reflexive use, as can be seen in the verbs like **vaspitaváti** (to raise, educate), **vézati** (to tie)**, vijòriti** (to fly (a flag)) etc. Impersonal reflexive verbs are also presented under number II. Within the headword, the particle **se** is placed in round brackets if the verb has the same meaning with or without it, e.g., **vijùgati (se)** (to wind).

As with the other mentioned Serbo-Croatian dictionaries, the DMNLL dictionary does not provide explicit data about the (in)transitivity of verbs since this is included in the definitions and examples. Only occasionally, when it conditions the realization of a certain meaning, is direct data given about the government or the right valency of verb lexemes and, more often, sub-lexemes, for example, **àdaknuti** (someone) *to expel, to drive away*; **bògatiti** I (someone, something) *to make rich, to enrich materially or spiritually*; **adresírati 2.** (to someone) *to intend/aim, usually a criticism or objection*; **vjenčávati 3.** (with something) *to adorn, to decorate*, etc. As seen from the provided examples, complements are listed in the form of indefinite pronouns for people and things in the appropriate case, placed in round brackets before the definition. However, they can also be found within the definition, e.g., **bŕziti 1.** *to hurry, to urge (someone) to move or work faster.* Regarding the left valency, only occasionally, after the data on aspect and forms, is data provided about the logical subject, e.g., with **bàstati** (with the logical subject in dative, rarely in accusative), but it is absent, for example, with **bòljeti,** although this verb in every mentioned meaning is realized in a construction with the accusative of the logical subject.

As with nouns and adjectives, attention is occasionally given to the collocation of verbs. Predominantly, this concerns their propensity for a certain form or combination with specific types of words, for instance, **brȉsāti 5.** (usually in imperative) *to disappear, to flee, to move away*; **bèndati** (usually with negation) *to consider, to heed, to attach importance, to notice; to respect, to care, to worry*; **blagovòljeti** (usually with the infinitive) *to honor someone with something, to express a willingness to do something, to condescend (in expressing respect, sometimes ironically)*, etc. In rarer instances, a particular semantic realization may be conditioned by a specific form, and this is also noted, for example, **vȉdjeti 10. a.** (in the 2nd person singular imperative) *in an indefinite meaning, when expressing astonishment, surprise, warning, threat, reproach*. **b.** (in the 2nd person singular present) *when emphasizing the content of the statement*, etc.

4.5 Adverbs, prepositions, conjunctions, interjections and particles

In addition to adverbs, prepositions, conjunctions, and interjections, corresponding abbreviations are used as part of speech markers (e.g., "*adv.*" for adverb, "*prep.*" for preposition, "*conj.*" for conjunction, "*interj.*" for interjection), for example:

— **avanturìstički** adv. *meaning in an adventurous manner, or in the spirit of an adventurer;*

— **van** prep. (with genitive) indicating **1.** *outside the boundaries of a space; beyond the framework of something.* **2.** *exceeding boundaries and frameworks, not adhering to them:* **a.** *above, over.* **b.** *beside, against.* **3.** *archaic for excluding: except, besides;*

— **a** conj. meaning **1.** *adversative:* **a.** *to connect independent sentences in a complex one.* **b.** *to link words that are in direct opposition: old yet naive.* **2.** *cumulative:* **a.** *to connect independent sentences in a complex sentence:* and. **b.** *to link words in a parallel relationship: thin but tall.* **3.** *disjunctive, to connect sentence parts of opposite meaning from which one must choose:* or. **4.** (when emphasized: ȁ) *traditionally temporal: when; as soon as, just;*

— **âj** interj. *is used to express various emotional states — most often pain, sorrow,* etc.

Generally, as seen from the examples provided, their meanings and usage domains are elaborately given. For prepositions, the cases they are used with are always indicated in round brackets. With particles, the full name of the word type is provided within the definition, e.g., **vàljda** *particle for indicating or highlighting hope, probably.*

Since adverbs of manner can be compared, their comparative forms are only listed if they exhibit accentual deviations and phonetic alternations, for example, **bȓzo** [comparative bȑže/bȑžē] (fast). Suppletive comparative forms have separate dictionary entries, for instance, **bȍlje** and **bȍljē** 1. *comparative of good (dobro) ...*

Other functions of invariable word types are also noted, such as the adverb **alègro 1.** *quickly, cheerfully, lively.* **2.** (in a nominal function) masculine, *a fast musical movement, a piece performed quickly*; the interjection **bȁmbajāt 1.** *upon falling.* **2.** (in the function of an invariable adjective) *at death's door, barely alive; dead,* etc.

5. Conclusion

The lexicographic marking of the grammatical features of lexemes is one of the more significant issues in both theoretical and practical lexicography. It is particularly crucial for a general dictionary of a language, as it is essential to both describe and prescribe the language. The type, scope, and manner of presenting grammatical data primarily depend on whether the dictionary is in electronic or print form. In the case of an electronic dictionary, the data should ideally be as comprehensive and explicit as possible, enhancing its informative power. In contrast, for a print dictionary, the presentation of grammatical data tends to be more economical and, therefore, often indirect due to space constraints. Consequently, it requires a higher level of lexicographic competence from potential users.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1908 (Article)

The DMNLL began as a print version, although the identification of sources and the extraction of lexemes were performed using modern computer programs. Therefore, its authors had to be mindful of space-saving and the economy of the lexicographical description. As for the grammatical treatment of headwords, this implies the mandatory indication of the word type, either directly, with a special marker, or indirectly with a marker of some comprehensive grammatical category or through the definition. The first case involves invariable parts of speech (adverbs, adjectives, conjunctions, interjections, and particles) and nominal pronouns. The second case pertains to nouns, adjectives, and adjectival pronouns, which are consistently marked by grammatical gender, and verbs, which receive obligatory aspect notation. Other types of grammatical data including categories of case and number, comparative forms of adjectives, verbal reflexivity, valency, specific functions, and collocational preferences of headwords are generally provided only if they are systematically unpredictable or if the meaning of the lexeme depends on them. Only adjectives that distinguish aspect have the form of the definite aspect indicated, although from an economy standpoint, this is unnecessary because it is predictable.

The described approach reflects a considerable reliance on traditional Serbo-Croatian lexicography, which is expected given the developmental trajectory of the Montenegrin language standard and Montenegrin studies. However, on the one hand, there was a missed opportunity to correct some omissions in the dictionary of the Serbo-Croatian language already identified in the literature, such as the non-marking of singularia tantum nouns (Ristić 2003: 128). Of course, since certain mass and abstract nouns, as uncountable, can also have plural forms beyond their basic meaning, when they are associated with e.g. something concrete, this should also be noted in their lexicographical description. We also consider the omission of noun valency to be an inherited shortcoming that should be corrected in future work. On the other hand, nominal case forms are grouped within the grammatical block in square brackets, as are verb forms, which seems to be a better solution than traditional. A useful innovation is the marking of the natural gender of nouns when it does not align with the grammatical gender.

Despite the clear intention in DMNLL to consistently provide grammatical features, the systematic nature is occasionally disrupted due to the lexicographers' oversight, which we particularly indicated in our analysis. Nevertheless, it can be stated that, in terms of grammatical data presented through transparent metalanguage, it is characterized by a high degree of consistency in the compilation of dictionary entries, and, along with removing observed omissions, it represents a solid foundation for the continuation of a serious and necessary lexicographic project such as the creation of a general dictionary of the Montenegrin language.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1908 (Article)

Endnotes

1. In 2006, Montenegro restored its independence and the following year proclaimed Montenegrin as the official language. However, since Montenegro is home to not only Montenegrins but also members of the Serbian, Bosniak, and Croatian peoples, Serbian, Bosnian, and Croatian languages are also in official use, being very close linguistic standards that emerged through a kind of layering of Serbo-Croatian after the dissolution of the Yugoslav state, in whose different versions Montenegro had existed since 1918.
2. As a thesaurus and academic dictionary, DSCLNL is one of the most important and demanding projects of the Serbian Academy of Sciences and Arts. The first volume was published in 1959, and to date, 21 volumes have been released. Once completed, it will contain over 35 volumes with more than 500,000 headwords. Although the sources and methodology of work are presented in the introduction of the first volume, it is entirely expected that they have been supplemented and changed during its many decades of compilation.
3. DSCLL is a six-volume dictionary, with 150 000 entries. The first three volumes were published in 1967 and 1969 in collaboration between Matica srpska and Matica hrvatska, while the latter three were independently published by Matica srpska in 1971, 1973, and 1976 after the Croatian side withdrew due to disagreements over the dictionary's concept and the nature of the common literary language.
4. Dictionaries, even if they claim to be solely descriptive or prescriptive, generally combine these two approaches because by describing the lexicon, they are essentially describing the linguistic norm, and the public expects from them at least a certain degree of normativity (Vrbinc et al. 2020: 576).
5. Alongside the headwords that we take as examples, from the dictionary entry we cite data only about those features that we wish to illustrate, and we do so in the manner presented in the DMNLL. In addition to the original language, headwords are provided in parentheses in English whenever possible, while their definitions are always translated.
6. Hoey (in Atkins and Rundell 2008: 304-305) views collocation as a "midway relation between grammar and collocation," explains that, for instance, a countable noun that almost always appears in the plural and never at the beginning of a sentence represents "a prima facie case of colligation — an observable preference for a subset of the available grammatical options." Besides nouns, they discuss the colligational tendencies of verbs and adjectives.

References

A. Dictionaries

DMNLL (*Dictionary of Montenegrin National and Literary Language*). 2016. *Rječnik crnogorskog narodnog i književnog jezika* I–. Podgorica: CANU.

DSCLNL (*Dictionary of Serbo-Croatian Literary and National Language*). 1959–(2020). *Rečnik srpskohrvatskog književnog i narodnog jezika* I–(XXI). Beograd: SANU.

DSCLL (*Dictionary of Serbo-Croatian Literary Language*). 1967–1976. *Rečnik srpskohrvatskoga književnog jezika* 1–6. Novi Sad: Matica srpska (1–3, Zagreb: Matica hrvatska).

B. Literature

Apresjan, J.D. 2010. *Prospekt aktivnogo slovarja russkogo jazyka*. Moskva: Jazyki slavjanskih kultur.

Atkins B.T. and M. Rundell. 2008. *The Oxford Guide to Practical Lexicography.* Oxford/New York: Oxford University Press.

De Caluwe, J. and A. van Santen. 2003. Phonological, Morphological and Syntactic Specifications in Monolingual Dictionaries. Van Sterkenburg, Piet (Ed.). 2003. *A Practical Guide to Lexicography*: 71-82. Amsterdam/Philadelphia: John Benjamins.

Čirgić, A., I. Pranjković and J. Silić. 2010. *Gramatika crnogorskoga jezika.* Podgorica: Ministarstvo prosvjete i nauke.

Gortan-Premk, D. 1980. O gramatičkoj informaciji i semantičkoj identifikaciji u velikom opisnom rečniku. *Naš jezik* 24(3): 107-114.

Kačić, M. 1994. Rječnik i gramatika. *Filologija* 22–23: 297-302.

Katičić, R. 1994. Leksikografija i gramatika. *Filologija* 22–23: 281-286.

Kostić-Golubičić, M. 1997. Gramatika i rečnik: teorija i praksa, tradicionalno i netradicionalno. *Naučni sastanak slavista u Vukove dane* 26(2): 457-463.

Kostić-Tomović, J. 2017. *Savremena nemačka leksikografija.* Beograd: FOKUS — Forum za interkulturnu komunikaciju.

Kozyrev, V.A. and V.D. Černjak. 2015. *Leksikografija russkogo jazyka. Vek nynešnij i vek minuvšij.* Sankt-Peterburg: Izdatelstvo RGPU im. A.I. Gercena.

Marković, M.A. 2014. Gramatika u srpskim rečnicima. Dragićević, R. (Ed.). 2014. *Savremena srpska leksikografija u teoriji i praksi*: 69-91. Beograd: Filološki fakultet Univerziteta u Beogradu.

Popović, Lj. 2003. Integralni rečenični modeli i njihov značaj za lingvistički opis i analizu korpusa. *Naučni sastanak slavista u Vukove dane* 31(1): 201-220.

Ristić, S. 2003. Leksikografski metajezik i srpska deskriptivna leksikografija. *Naučni sastanak slavista u Vukove dane* 31(1): 119-130.

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: 68-89.

Tafra, B. 2005. *Od riječi do rječnika*. Zagreb: Školska knjiga.

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.

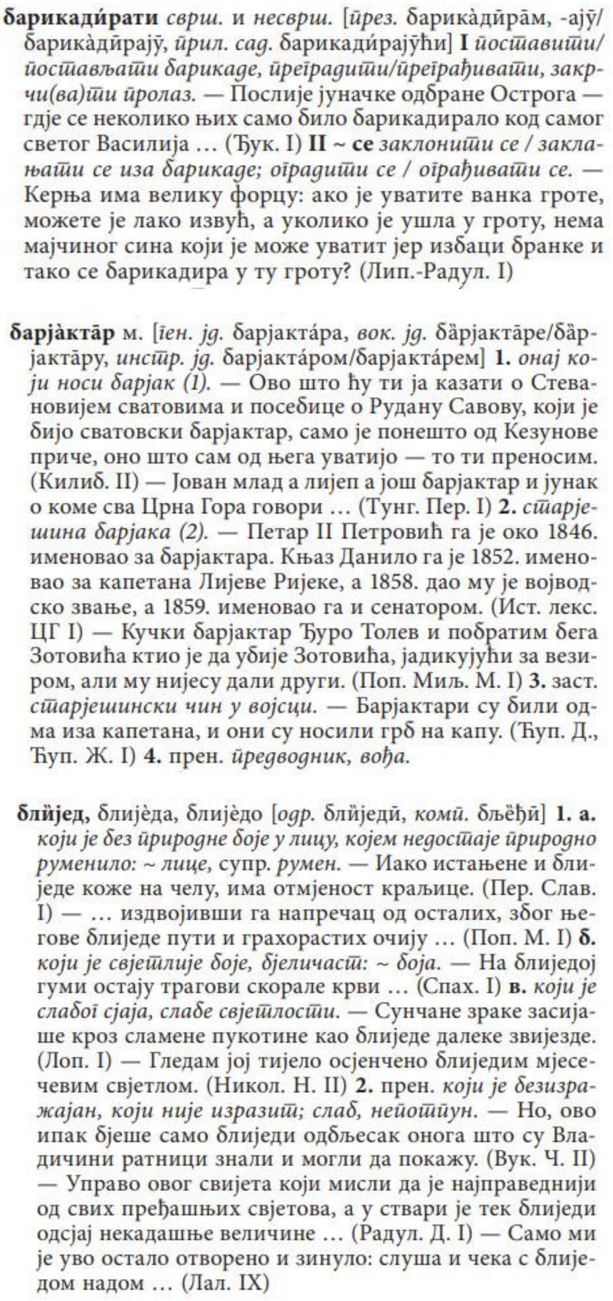
Vrbinc, M., D. Šipka and A. Vrbinc. 2020. Normative Labels in Two Lexicographic Traditions: A Slovene–English Case Study. *Lexikos* 30: 561-582.

Zgusta, L. 1991. *Priručnik leksikografije*. Sarajevo: Svjetlost.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1908 (Article)

Appendix A

Examples of the dictionary articles *barikadirati*, *barjaktar* and *blijed* in the DMNLL.



барикади́рати *сврш.* и *несврш.* [*през.* барикàдӣрāм, -ајӯ/барикàдӣрајӯ, *прил. сад.* барикади́рајӯћи] **I** *поставити/постављати барикаде, преградити/преграђивати, закрчи(ва)ти пролаз.* — Послије јуначке одбране Острога — гдје се неколико њих само било барикадирало код самог светог Василија … (Ђук. I) **II ~ се** *заклонити се / заклањати се иза барикаде; оградити се / ограђивати се.* — Керња има велику форцу: ако је уватите ванка гроте, можете је лако извућ, а уколико је ушла у гроту, нема мајчиног сина који је може уватит јер избаци бранке и тако се барикадира у ту гроту? (Лип.-Радул. I)

барјàктāр м. [*ген. јд.* барјактáра, *вок. јд.* бȁрјактāре/бȁрјактāру, *инстр. јд.* барјактáром/барјактáрем] **1.** *онај који носи барјак (1).* — Ово што ћу ти ја казати о Стевановијем сватовима и посебице о Рудану Савову, који је бијо сватовски барјактар, само је понешто од Кезунове приче, оно што сам од њега уватијо — то ти преносим. (Килиб. II) — Јован млад а лијеп а још барјактар и јунак о коме сва Црна Гора говори … (Тунг. Пер. I) **2.** *старјешина барјака (2).* — Петар II Петровић га је око 1846. именовао за барјактара. Књаз Данило га је 1852. именовао за капетана Лијеве Ријеке, а 1858. дао му је војводско звање, а 1859. именовао га и сенатором. (Ист. лекс. ЦГ I) — Кучки барјактар Ђуро Толев и побратим бега Зотовића ктио је да убије Зотовића, јадикујући за везиром, али му нијесу дали други. (Поп. Миљ. М. I) **3.** заст. *старјешински чин у војсци.* — Барјактари су били одма иза капетана, и они су носили грб на капу. (Ћуп. Д., Ћуп. Ж. I) **4.** прен. *предводник, вођа.*

бли̏јед, блиjѐда, блиjѐдо [*одр.* бли̏једӣ, *комп.* бље̏ђӣ] **1. а.** *који је без природне боје у лицу, којем недостаје природно руменило: ~ лице,* супр. *румен.* — Иако истањене и блиједе коже на челу, има отмјеност краљице. (Пер. Слав. I) — … издвојивши га напречац од осталих, због његове блиједе пути и грахорастих очију … (Поп. М. I) **б.** *који је свјетлије боје, бјеличаст: ~ боја.* — На блиједој гуми остају трагови скорале крви … (Спах. I) **в.** *који је слабог сјаја, слабе свјетлости.* — Сунчане зраке засијаше кроз сламене пукотине као блиједе далеке звијезде. (Лоп. I) — Гледам јој тијело осјенчено блиједим мјесечевим свјетлом. (Никол. Н. II) **2.** прен. *који је безизражајан, који није изразит; слаб, непотпун.* — Но, ово ипак бјеше само блиједи одбљесак онога што су Владичини ратници знали и могли да покажу. (Вук. Ч. II) — Управо овог свијета који мисли да је најправеднији од свих пређашњих свјетова, а у ствари је тек блиједи одсјај некадашње величине … (Радул. Д. I) — Само ми је уво остало отворено и зинуло: слуша и чека с блиједом надом … (Лал. IX)

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1912 (Lexiconote)

Defining Feminine Personal Nouns in Polish: A Practical and Postulative Overview Based on the *Dictionary of Polish Female Nouns*

Agnieszka Małocha, *Institute of Polish Philology,*
Philological Faculty, University of Wrocław, Wrocław, Poland
(agnieszka.malocha@uwr.edu.pl) (https://orcid.org/0000-0002-8698-2812)

Abstract: This article considers and analyses various ways of defining feminine personal nouns in Polish-language lexicography. It deals with the techniques of defining products of the feminitive word-formation category which have been recorded in historical and modern dictionaries. It aims to determine the best way to define feminine personal nouns from the perspective of what the modern user needs and how they perceive the world. Against the recognised definition practices, a proprietary method developed at the Formation of Feminine Personal Nouns Research Section at the Institute of Polish Studies of Wrocław University, Poland, has been proposed. It has become the basis for lexicographic description used in *Słownik nazw żeńskich polszczyzny* (Dictionary of Polish Female Nouns). This article puts forward a rationale for defining lexical items in which a reference to the generic masculine, i.e. a systemic motivation typical of presentations of Polish word formation, has been abandoned.

Keywords: FEMININE PERSONAL NOUNS, FEMININITIVE, POLISH WORD FORMATION, DICTIONARY, DEFINITION

Opsomming: Die definiëring van vroulike persoonsname in Pools: 'n Praktiese en voorveronderstelde oorsig gebaseer op die *Woordeboek van Poolse vroulike naamwoorde*. In hierdie artikel word verskeie metodes vir die definiëring van vroulike persoonsname in die Poolse leksikografie beskou en geanaliseer. Aandag word geskenk aan die tegnieke waarmee die resultate van die feminitiewe woordvormingskategorie wat in historiese en moderne woordeboeke opgeneem is, gedefinieer word. Dit het ten doel die bepaling van die beste metode om vroulike persoonsname vanuit die perspektief van die moderne gebruiker se behoeftes en beskouings van die wêreld, te definieer. Strydig met die erkende definiëringspraktyke, is 'n vertroulike metode wat by die Navorsingsafdeling vir die Vorming van Vroulike Persoonsname by die Instituut van Poolse Studies aan die Wrocław Universiteit, Pole, ontwikkel is, voorgestel. Dit het die grondslag vir die leksikografiese beskrywing wat in *Słownik nazw żeńskich polszczyzny* (Woordeboek van Poolse vroulike naamwoorde) gebruik word, gevorm. In hierdie artikel word 'n goed-gefundeerde voorstel vir die definiëring van leksikale items uiteengesit waarin afgesien is van 'n verwysing na die generiese manlike, m.a.w. 'n sistemiese motivering tipies van die aanbieding van Poolse woordvorming.

Sleutelwoorde: VROULIKE PERSOONSNAME, FEMINITIEF, POOLSE WOORDVORMING, WOORDEBOEK, DEFINISIE

1. Introduction

In Poland, feminitive derivatives (also referred to as feminitives) are entangled in history, and, consequently, have been a subject of lively normative, linguistic, political, ideological, identity-related and numerous other discussions for over a century. Małocha-Krupa (2018a, 2018b) defines feminitive derivatives as language forms that are synthetic (one-word) names of females. These are nouns containing a morphological (suffixal or paradigmatic) indicator of femininity; they designate professions, titles, academic degrees, social positions and attributes of women, e.g. *reporterka* [a woman reporter], *nauczycielka* [a woman teacher], *prezeska* [a woman president], *doktorka/doktora* [a woman doctor].

At present there is an increase in the number of feminine personal nouns in Polish. For many years feminine personal nouns have been the subject of animated — frequently emotional — discussions which are not always related to linguistic parameters (Łaziński 2006). Currently there are discussions taking place regarding the stylistic value of many feminine personal nouns, their acceptability in standard Polish, and shortcomings in their codification (Krysiak 2020, Szpyra-Kozłowska 2021), as well as the way in which they should be treated in dictionaries of general Polish (Małocha-Krupa 2018b, Śleziak 2018). To contemporary lexicographers they are not easy to codify, because, for example, of the difficulty of determining which of them are used in general Polish, and which are only limited to specific speech communities or (equality-promoting) contexts. However, we are steadily observing the transfer from these limited contexts or specific linguistic communities to general, unmarked language. Such stylistic peregrination and reclassification of feminitives still means that a contemporary lexicographer faces difficulties when attempting to objectively classify a given linguistic unit as belonging to e.g. a specific register. As Szpyra-Kozłowska (2021) notes, something that was marked stylistically only recently may have since become neutral.

A further problem for Polish lexicographers is related to the various traditions of defining feminine personal nouns as these traditions are mostly rooted in historical lexicographic practices (cf. Małocha-Krupa 2021). This problem was initially encountered by the team of female linguisticians from Wrocław University (Marta Śleziak, Patrycja Krysiak, Katarzyna Hołojda-Mikulska and Agnieszka Małocha), who compiled the *Słownik nazw żeńskich polszczyzny* (SNŻP — Dictionary of Polish Female Nouns; cf. Małocha-Krupa 2021). Currently, because of the problematic defining practices and the growing textual frequency of feminine personal nouns, as well as increased naming needs, lexicographic work is under way on a new, expanded edition of SNŻP, to be published in 2025.

2. What's new in the SNŻP II?

The first edition of the SNŻP contained 2,100 entries. The currently prepared second edition has twice as many entries. The lexicographers decided to extend their area of research and to register, apart from the contemporary, frequently innovative feminitives, such as: *boomerka* [a female boomer], *celebrytka* [a female celebrity], *couchsurferka* [a female couchsurfer], *genderystka* [a female gender researcher], *hipsterka* [a female hipster], *homofobka* [a female homophobe], *japiszonka/japówka* [a female yuppie], *singielka* [a female single], to words that are used to describe professions, customs, etc. that are now obsolete, e.g. *grabarka* [a gravedigger's wife or daughter], *pończoszniczka/pończoszarka* [a woman engaged in the making, repair or sale of stockings]. To that end, the task of excerpting material from historic sources was undertaken, starting from the oldest lexicographic monuments, such as *Dykcjonarz Jana Murmeliusza* (*Dictionarius Ioannis Murmellii variarum rerum*) from 1526. Material was also collected from other historic dictionaries such as *Słownik języka polskiego* [Dictionary of the Polish Language] by Samuel Linde (1807–1814), *Słownik wileński* [The Vilnius Dictionary] (1861) and *Słownik wrszawski* [The Warsaw Dictionary] (1900–1927), as well as from corpora containing historic material, such as the *Elektroniczny korpus tekstów polskich z XVII i XVIII w* [Electronic Corpus of 17th- and 18th-century Polish Texts — https://korba.edu.pl/query_corpus/].

To date, about 5,000 personal nouns have been collected (as of 15 April 2024). The collection, which not only enriches the knowledge of contemporary linguistics on the productivity of the very feminine personal noun category, but also provides many valuable examples of semantic changes that have occurred in Polish, e.g. as regards the evolution of the semantic category of possessiveness or within homonymic relations.

3. How to define feminitives in Polish

The proposed practice for explicating meaning outlined here, developed and adopted by the lexicographical team preparing the SNŻP II, is in response to some opinions expressed after the publication of the first edition of the SNŻP, especially those by Łaziński (2016: 127): "The editors seem not to have agreed on how the definitions and genus proximum, i.e. the root personal name, should look like."

The editors of the SNŻP II, namely Katarzyna Hołojda-Mikulska, Patrycja Krysiak, Agnieszka Małocha, Simone Pasternak, Marta Śleziak, Julia Wójcik, Małgorzata Winnicka and Alicja Wrzyszcz, faced, as suggested, the lexicographic problem of how to define feminine personal nouns. So far, Polish lexicography has been dominated by models of explicating meaning by reference to the generic masculine form. It was recognised that a definition using the masculine form was the optimum one, as it was in agreement with the structural, systemic thinking about language. In Polish, more often than not, feminine personal nouns derive from masculine nouns. It was therefore accepted that the most appropriate way of defining them was to refer to the relevant masculine root of a given feminine personal noun, i.e. — within that meaning (a feminine personal noun as a modificational derivative of the masculine form) — *nauczycielka* used to be defined as a 'woman teacher'. A similar model for explication of meanings dominated in numerous Polish dictionaries, i.e. definitions made references to systemic Polish word-formation knowledge, e.g.

— *adwokatka* was defined as 'a female advocate'
— *autorka* was defined as 'a female author'
— *wariatka* was defined as 'a female lunatic'

Some dictionaries also explicate meaning by referencing the derived feminitive to the equivalent of the masculine personal noun, e.g.

— *nauczycielka* — a feminine form of the noun *nauczyciel* [a male teacher]
— *autorka* — a feminine form of the noun *autor* [a male author]
— *realistka* — a feminine form of the noun *realista* [a male realist]
— *zabójczyni* — a feminine form of the noun *zabójca* [a male killer]

The editors of the SNŻP decided not to define meanings by means of the generic masculine form. They recognised that to a modern recipient, systemic (structural) knowledge is of secondary importance in the light of 'naive, common-sense conceptualisation' (Apresjan 1980: 79-84, Anusiewicz 1992: 9-11). Research has shown (e.g. Tambor 2013) that to many users of contemporary Polish, the word *nauczycielka* means 'a woman that teaches/trains,' not 'a female teacher.' Thus, due to the nature of the registered material, the systemic method of explicating word-formation roots and using them to describe meanings has been abandoned in the SNŻP. The decision was not entirely innovative or trailblazing, as such a description of linguistic units was already applied in some historical dictionaries, e.g. *Słownik wileński* [The Vilnius Dictionary] which describes *nauczycielka* as 'a female who teaches someone' (and also: '*zbawczyni* [a female saviour], *oswobodzicielka* [a female liberator]').

The editors of *Wielki Słownik Języka Polskiego* [The Great Dictionary of the Polish Language] also did away with using the masculine form to define meaning and included information about the word-forming motivation of feminine derivatives in the tab *Pochodzenie* [Origin]. By way of example, the innovative entry *forumowiczka* is defined as 'a woman expressing herself on an online forum' (instead of the systemic: '*kobieta forumowicz*' [a woman forum member] — all the more so that the noun *forum*, not the masculine form *forumowicz* [a man that participates in a forum], was rightly recognised as the word-formation root.

The way of defining meanings used in the SNŻP reflects the simplest text conceptualisation model, corresponding to naive, everyday, common-sense knowledge about the defined entry. Description of feminine personal nouns by referring to the generic masculine form is a construct typical of inside-system thinking, incompatible both with an equality-based vision of the world, linguistic symmetry theories or a cultural way of thinking about a female subject expressed by means of a specific morphological form. The speaking subject does not perceive a woman, e.g. *nauczycielka* [a woman that teaches] in terms of a sum of semantic components [+ femininity *woman* + the generic masculine form *nauczyciel*]. Thus, for the SNŻP, a decision was made not to update the systemic motivation model, which is very useful and valuable elsewhere, e.g. in the theory of descriptive word formation, in establishing formal and structural relations that occur between a derivative and its root, and to define meanings in a way that omits relations between feminine nouns and masculine nouns/the generic masculine or other possible word-formation motivations (Skarżyński 1999: 64).

Consequently, such an approach facilitated resolution of the problem of how to explain meanings of feminine personal nouns with meaning asymmetry in relation to their masculine roots, both from the (1) diachronic, and (2) synchronic perspectives. Semantic asymmetry is illustrated for instance by a pair of extant nouns (1): *spodniarz* and *spodniarka. Spodniarz* means 'a man sewer specialising in sewing trousers', but what about *spodniarka*? It may be assumed that it was understood symmetrically as 'a female sewer specialising in sewing trousers.' However, because of the fact that culturally and socially, in the 19th century, it was not acceptable for a woman to sew trousers for men, the historical dictionaries define *spodniarka* pejoratively: 'an annoying and cocky woman with a lower social status'. Differentiation of the referential potential and semantic scope between contemporary masculine and feminine forms is also quite common. This is exemplified by (2) the pair of nouns *ambasador* and *ambasadorka*. The generic noun *ambasador* refers most frequently to 'a person who is a highest-ranking diplomatic representative of a state in relations with the authorities of another state or international organisation', while the feminine personal noun *ambasadorka* predominates in social custom with a more modest-ranking reference. It means 'a woman promoting an issue or idea or protecting someone's interests'. Therefore, it is clear that defining a feminine personal noun by referring to its corresponding masculine noun is not always a fail-safe method of defining feminine nouns. The decision by the editors of the SNŻP not to define feminine personal nouns by referring to the generic masculine has also simplified explication of meanings of nouns that do not derive from the masculine, e.g. *cygarniczka* once meaning 'a woman that was a manual worker in a cigar manufacturing plant in Kraków'; the word's masculine equivalent has never been used.

The history of feminine personal nouns has numerous lexical and semantic asymmetries which can be shown if definitions are based on semantic relations and the convention of semantic interpretation is adopted, while formal interpretation is used in structural descriptions. Especially considering the fact that in word-formation theory, establishment of relations between the derivative and its root may be characterised by the existence of numerous motivations (Skarżyński 1999: 64). Ultimately, the editors of the SNŻP decided not to refer to the word-formation motivation of the analysed feminine personal nouns when explaining their meanings, and to use the following forms as the initial element of the definiens:

(1) the demonstrative pronoun *ta* ['that' referring to a female/feminine noun]:

- — *abstynentka* is defined as 'that [female] who gives up on something, especially drinking alcohol',
- — *administratorka* is defined as 'that [female] who administers, runs something';

(2) or other forms resulting from the semantic relations in which a given personal noun functions. Thus, some definitions with *genus proximum* are used, e.g. *pracowniczka* [that [female] who works], *specjalistka* [that [female] who is a specialist in something], *wykonawczyni* [that [female] who does something], *zawodniczka* [that [female] who competes], *znawczyni* [that [female] who knows something very well], *zwolenniczka* [that [female] who supports something or someone], for instance:

- — *aptekarka* is 'a woman who works in or owns a pharmacy'; *farmaceutka*[1]
- — *archeolożka* is 'a woman specialising in archaeology'
- — *kaskaderka* is 'a woman performing an especially risky and dangerous movie scene, usually instead of an actor or actress'; *dublerka; zastępczyni*
- — *chodziarka* is 'a woman competitor in racewalking'
- — *fashionistka* is 'a woman expert in fashion, a woman fashion designer'
- — *demokratka* is 'a woman supporter of the democratic system, a woman with democratic persuasion'

Furthermore, some definitions are encyclopaedic in nature. In particular, this applies to feminine personal nouns that today are not very transparent semantically, as they refer to concepts or ideas that are less known. Examples include:

- — *Katofeministka* is a Catholic woman who supports some feminist views, especially those connected with Christian feminism, which recognises the Church and its teachings, but which also aims to raise women's status in the Church and to redefine the grounds for theological tradition, regarded as androcentric from the feminist perspective.
- — *Fenomenolożka* is a woman supporter, representative of phenomenology — an ontological direction in philosophy that assumes the possibility of reaching an object thanks to its direct observation; phenomenology was formulated by Edmund Husserl and developed by Roman Ingarden, who adapted inspirations of the direction to the philosophy of art, aesthetics and theory of literary work and its cognition; Ingarden's key terms introduced into literary theory include: the diegetic world, places of indeterminacy, concretization and the idea of ideation, which is relevant to the entire current of phenomenology and which is connected with a reduction in incidental (inessential) features of an object.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1912 (Lexiconote)

— *Flapperka* in the 1920s was a young woman from the West who opposed the then prevailing canon of femininity, who had an independent, ostentatious and provoking lifestyle, who wore clothes masking her womanly shape and short hair; *chłopczyca*; *chłopobaba*.

Regular, systematic use of the defining techniques referred to above will enable presentation of the collected lexical material of the SNŻP in a clear way, in line with the lexicographical method adopted by the dictionary editors.

4. Conclusions

The SNŻP is the result of many years of work by a team of experienced Polish linguists that are keenly aware of the shifting linguistic landscape, especially as it pertains to how modern Polish reality is reflected in the language. This reality has been changing and transforming very dynamically in recent decades. In contemporary Polish texts we observe increased productivity, or even an expansion, of female nouns. It seems understandable, therefore, that new tasks are emerging for lexicography: the codification of innovative feminitives, but also the revival of old ones, which nowadays often undergo neosemantization and various semantic evolutions.

The new edition of the SNŻP, currently under preparation, requires many new, innovative solutions in defining/describing feminine lexical items. As noted, the issues on defining feminine lexical items are complex and rooted in historical lexicographic or linguistic traditions. This is supported by the current linguistic discourse in Poland and with the publication of the first edition of the SNŻP, the difficulties with defining feminitives again became apparent, especially as these definitions did not necessarily meet the needs of the contemporary dictionary user. Therefore in the forthcoming SNŻP, defining through reference to grammatical and word-formation knowledge, that is, to the masculine form (which is often the word-formation base of a given feminitive), has been abandoned. This decision is based on the assumption that the description of feminine personal nouns by referring to the generic masculine form is a construct typical of inside-system thinking, incompatible both with an equality-based vision of the world, linguistic symmetry theories or a cultural way of thinking about a female subject expressed by means of a specific morphological form. Therefore, a method of explicating meaning was chosen that reflects the simplest text conceptualisation model, corresponding to naive, everyday, commonsense knowledge about the defined entry. Sometimes it was deemed necessary to include encyclopedic definitions, especially for lexemes that are less recognizable and less semantically transparent. The choice of these methods of defining seems to be in line with the needs of contemporary users, and it also allows for consistency and clarity in the presentation of lexical material.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1912 (Lexiconote)

Endnote

1. In some entries, the definition ending with a semicolon is followed by synonyms of the given lemma.

References

Anusiewicz, J. 1992. *Potoczność jako sposób doświadczania świata i jako postawa wobec świata* [Informality as a Way of Experiencing the World and as an Approach to the World]. Anusiewicz, J. and F. Nieckula (Eds.). 1992. *Język a Kultura* 5: 9-20. Wrocław: Wiedza o Kulturze.

Apresjan, J. 1980. *Semantyka leksykalna. Synonimiczne środki języka* [Lexical Semantics. Synonymous Means of Language — Translated by Kozłowska, Z. and A. Markowski]. Wrocław: Zakład Narodowy im. Ossolińskich.

Elektroniczny korpus tekstów polskich z XVII i XVIII w [Electronic Corpus of 17th- and 18th-century Polish Texts].
https://korba.edu.pl/query_corpus/ [9 April 2024]

Krysiak, P. 2020. *Nazwy żeńskie we współczesnej leksykografii polskiej i francuskiej. Analiza porównawcza wybranych najnowszych polskich i francuskich słowników ogólnych* [Feminine Personal Nouns in Contemporary Polish and French Lexicography. Comparative Analysis of Selected Most Recent Polish and French General Dictionaries]. Wrocław: Oficyna Wydawnicza ATUT.

Łaziński, M. 2006. *O panach i paniach. Polskie rzeczowniki tytularne i ich asymetria rodzajowo-płciowa* [On Men and Women. Polish Titular Nouns and Their Sex/Gender Asymmetry]. Warsaw: Wydawnictwo Naukowe PWN.

Łaziński, M. 2016. Review: Małocha-Krupa, Agnieszka (Ed.). 2015. *Słownik nazw żeńskich polszczyzny* [Dictionary of Polish Feminine Personal Nouns]. *Język Polski* XCVI(3): 125-128.

Linde, S.B. 1807–1814. *Słownik języka polskiego* [Dictionary of the Polish Language]. Vols. 1–6. Warsaw.

Małocha-Krupa, A. (Ed.). 2015. *Słownik nazw żeńskich polszczyzny* [Dictionary of Polish Female Personal Nouns]. Entries by: Hołojda, K., P. Krysiak, A. Małocha-Krupa and M. Śleziak. Wrocław: Wydawnictwo Uniwersytetu Wrocławskiego. [SNŻP]

Małocha-Krupa, A. 2018a. *Feminatywum w uwikłaniach językowo-kulturowych* [Feminitives and Their Linguistic and Cultural Entanglements]. Wrocław: Oficyna Wydawnicza ATUT.

Małocha-Krupa, A. 2018b. Opis leksykograficzny feminatywum. (Nie)możliwości zobiektywizowania [Lexicographic Description of Feminine Personal Nouns. (In)ability to Objectivise]. Bańko, M. and H. Karaś (Eds.). 2018. *Między teorią a praktyką. Metody współczesnej leksykografii* [Between Theory and Practice. Methods in Contemporary Lexicography]: 151-165. Warsaw: Wydawnictwo Uniwersytetu Warszawskiego.

Małocha-Krupa, A. 2021. Feminine Personal Nouns in the Polish Language. Derivational and Lexicographical Issues. *Lexikos* 31: 101-118.
https://lexikos.journals.ac.za/pub/article/view/1630

Murmeliusz, J. 1526. *Dictionarius Ioannis Murmellii variarum rerum.* Kraków.

Skarżyński, M. 1999. *Powstanie i rozwój polskiego słowotwórstwa opisowego* [Origin and Development of Polish Descriptive Vocabulary]. Kraków: Wydawnictwo UNIVERSITAS TAiWPN.

Słownik języka polskiego (so-called *Słownik warszawski* [Dictionary of the Polish Language (also referred to as 'The Warsaw Dictionary')]). 1900–1927. Karłowicz, J., A.A. Kryński and W. Niedźwiedzki (Eds.). Vols. 1–8. Warsaw: Nakładem prenumeratorów i Kasy im. Mianowskiego [and others].

Słownik języka polskiego (so-called *Słownik wileński* [Dictionary of the Polish Language (also referred to as 'The Vilnius Dictionary')]). 1861. Zdanowicz, A. et al. (Eds.). Vols. 1–2. Wilno: M. Orgelbrand.

Szpyra-Kozłowska, J. 2021. *"Nianiek", "ministra" i "japonki". Eseje o języku i płci* ['Nianiek', 'ministra' and 'japonki'. Essays on Language and Gender]. Kraków: UNIVERSITAS.

Śleziak, M. 2018. Prymat leksykograficznego egzemplum i jego konsekwencje dla hasła — na przykładzie Słownika nazw żeńskich polszczyzny [Primacy of Lexicographical Example and Its Consequences for the Entries — On the Example of the Dictionary of Polish Feminitives]. Bańko, M. and H. Karaś (Eds.). 2018 *Między teorią a praktyką. Metody współczesnej leksykografii* [Between Theory and Practice. Methods in Contemporary Lexicography]: 245-255. Warsaw: Wydawnictwo Uniwersytetu Warszawskiego.

Tambor, J. 2013. *Kobiety w języku polskim. O formacjach żeńskich raz jeszcze z perspektywy języka polskiego jako obcego* [Women in the Polish Language. About Feminitive Formations Once Again from the Perspective of Polish as a Foreign Language]. Burzyńska-Kamieniecka, A. and A. Libura (Eds.). 2013. *Sapientia ars vivendi. Księga Jubileuszowa ofiarowana Profesor Annie Dąbrowskiej* [A Jubilee Book Dedicated to Prof. Anna Dąbrowska]: 345-359. Wrocław: Wydawnictwo ATUT.

Żmigrodzki, P. and M. Bańko (Eds.). *Wielki Słownik Języka Polskiego* [The Great Dictionary of the Polish Language].
https://wsjp.pl/ [15 April 2024]

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1914 (Article)

Adaptation and Validation of the Strategy Inventory for Electronic Dictionary Use (S.I.E.D.U.) for Chinese Learners

Lingling Li, *School of Literature and Media, Dongguan University of Technology, Dongguan, China*
(lill@dgut.edu.cn) (https://orcid.org/0000-0003-2069-0056)

Hui Wang, *Zhongshan Medical School, Sun Yat-sen University, Guangzhou, China*
(Corresponding Author, wangh@mail.sysu.edu.cn)
(https://orcid.org/0000-0002-2131-977X)
and
Hai Xu, *Centre for Linguistics and Applied Linguistics, Guangdong University of Foreign Studies, Guangzhou, China*
(xuhai1101@gdufs.edu.cn) (https://orcid.org/0000-0003-4644-9033)

Abstract: This study aimed to adapt and validate the S.I.E.D.U. questionnaire for assessing electronic dictionary (ED) use in the Chinese context. Six specialists translated the questionnaire into Chinese and then back-translated it into English to ensure accuracy. The Chinese version was administered to 518 participants. Factor analysis revealed seven factors, differing from the original four-factor structure. In addition to factors related to ED conventions, functions, and strategic skills, the study identified additional factors, including learners' preparation and troubleshooting, acceptance and usage context, storage format and advantages of ED, and ED subscription. These findings provide a comprehensive understanding of ED use strategies from the perspective of Chinese learners, benefiting both learners and educators. This study not only validates the S.I.E.D.U. in the Chinese context but also underscores the importance of enhancing learners' strategies for effectively utilizing electronic dictionaries.

Keywords: ELECTRONIC DICTIONARY, ELECTRONIC DICTIONARY USE QUESTIONNAIRE, ELECTRONIC DICTIONARY USE STRATEGIES, LANGUAGE LEARNERS, CHINESE CONTEXT, CULTURAL ADAPTATION

Opsomming: Die aanpassing en bevestiging van die Strategie-inventaris vir Elektroniese Woordeboekgebruik (S.I.E.W.G.) vir Chinese leerders. In hierdie studie is beoog om die S.I.E.W.G.-vraelys vir die assessering van elektroniese woordeboekgebruik (EW-gebruik) in die Chinese konteks aan te pas en te bevestig. Ses kundiges het die vraelys in Chinees vertaal en dit toe terugvertaal in Engels om die akkuraatheid daarvan te waarborg. Die Chinese weergawe is aan 518 deelnemers uitgedeel. 'n Faktoranalise het sewe faktore blootgelê wat van die oorspronklike vier-faktorstruktuur afwyk. Bykomend tot faktore wat verband hou met EW-konvensies, -funksies, en strategiese vaardighede, het die studie addisionele faktore, insluitende leerdersvoorbereiding en foutopsporing, aanvaarding en gebruikskonteks, bergingsformaat en voordele van EW, en EW-intekening, geïdentifiseer. Hierdie bevindings verskaf 'n omvattende begrip van EW-gebruikstrategieë vanuit die perspektief van Chinese leerders wat beide leerders en opvoeders baat. Dié studie bevestig nie net die S.I.E.W.G. in die Chinese konteks nie, maar beklemtoon ook die belangrikheid van die uitbreiding van leerderstrategieë vir die effektiewe benutting van elektroniese woordeboeke.

Sleutelwoorde: ELEKTRONIESE WOORDEBOEK, VRAELYSTE VIR ELEKTRONIESE WOORDEBOEKGEBRUIK, STRATEGIEË VIR ELEKTRONIESE WOORDEBOEKGEBRUIK, TAALLEERDERS, CHINESE KONTEKS, KULTURELE AANPASSING

1. Introduction

Dictionaries play a crucial role in the language learning process (Campoy-Cubillo 2021). According to Nation (2001: 296), "dictionary use is a kind of language-focused learning: the deliberate, explicit study of words". Specifically, they aid learners in their acquisition of essential information necessary for effective communication, both within and outside the classroom (McAlpine and Myles 2003). Serving as a dual-purpose instrument for comprehension and production, dictionaries empower learners to enhance their language skills. Therefore, learners are encouraged to integrate dictionary use with other learning strategies to fully utilize the wealth of information dictionaries provide (Summers 2013).

The evolution of information technology has led to the digitization and storage of dictionary entries in electronic formats. Defined as "any reference material stored in electronic form that gives information about spelling, meaning, or use of words" (Nesi 2000: 839), electronic dictionaries (EDs) have emerged as an alternative to their traditional paper counterparts. In contrast to the latter, EDs boast additional features, including audio pronunciation, video, animation, etc. (Nesi 2000; Joffe 2009; Winkler 2001). Moreover, EDs provide advanced look-up routes such as fuzzy searches, hyperlinks, pop-up windows, data boxes, multi-access, and menus (Gouws and Prinsloo 2021; Pastor and Alcina 2010; Rundell 2013; Verlinde et al. 2009). These functions not only enrich the language learning experience but also make the dictionary consultation more engaging and effective. Besides, the enhanced functionality and convenience offered by EDs have rendered them indispensable tools for language learning. Accordingly, learners are compelled to adapt and employ novel strategies, such as leveraging hyperlinks to explore related words and concepts, and utilizing fuzzy search functions to find words even with uncertain spelling, to effectively harness the potential of EDs in their language acquisition endeavors.

To the best of our knowledge, the Strategy Inventory for Electronic Dictionary Use (S.I.E.D.U.) (Mavrommatidou et al. 2019) represents the pioneering effort to assess Greek users' strategies in electronic dictionary (ED) searches. Mavrommatidou et al. (2019) emphasize the necessity of employing the S.I.E.D.U.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1914 (Article)

in subsequent research endeavors to bolster its reliability and validity. Furthermore, they call for its translation and cultural adaptation into other languages. However, there exists a notable gap in the availability of an instrument specifically tailored to gauge Chinese learners' strategies for ED use. Adaptation and validation of the S.I.E.D.U. to assess Chinese learners' ED strategies are imperative for several reasons. Firstly, while numerous studies have delved into ED use and reference skills (e.g. Campoy-Cubillo 2021; Klein 2008; Krajka 2015; Pastor and Alcina 2010), relatively few have examined ED strategies from learners' perspectives. Comparing the results of these studies with learners' perceptions of their lookup process poses a challenge. Learners may lack awareness of the strategies they employ or the strategies they need to improve their ED consultation practices. Therefore, understanding learners' perspectives on ED strategies is essential for enhancing their use of ED effectively. Secondly, the present study, focusing on the appropriate adaptation protocol, aims to optimize the reliability and validity of the S.I.E.D.U. in the Chinese context. Although the S.I.E.D.U. was initially tailored for Greek learners, this adaptation aims to assess the applicability of the S.I.E.D.U. in examining electronic dictionary use strategies (ED use strategies) not only among Greek learners but also for learners from other cultures, demonstrating its general utility. Additionally, it seeks to provide valuable insights for learners and educators, thereby enriching pedagogical lexicography in a broader context.

This study details the adaptation and validation of the S.I.E.D.U. questionnaire for use in the Chinese cultural context. By validating this adaptation, the study aims to not only demonstrate the utility of the questionnaire in effectively tapping into learners' strategies but also contribute to advancing understanding and research in ED use across different linguistic and cultural settings.

2. Literature review

As a language learning strategy, dictionary use, which falls into the category of self-regulation (e.g. O'Malley and Chamot 1990; Oxford 1990), is conducive to vocabulary acquisition (e.g. Fan 2003; Gu 2003; Li and Xu 2015). It is also a preferred strategy during the reading and writing process, with an impact on reading comprehension and writing proficiency (e.g. Cohen and Oxford 2002; Harvey and Yuill 1997).

Despite the considerable attention given to dictionary use as a strategy in various domains, there has been limited exploration of how strategic dictionary use activates appropriate skills in relevant contexts. Gavriilidou (2011, 2013) laid the groundwork for clarifying the concept of dictionary use strategies, connecting the descriptive notion of reference skills with the established theory of learning strategies (Oxford 1990). These strategies encompass the techniques employed by skilled dictionary users to conduct quick and effective searches. Dictionary use strategies are subsequently defined as "users' decisions and behaviors regarding the internal processes they adopt in order to perform suc-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1914 (Article)

cessful dictionary searches, the significance of which is prominent in the case of vocabulary acquisition, reading comprehension, and writing" (Chadjipapa et al. 2020: 444). These strategies are viewed as integral components of broader learning strategies, indicating that users' strategic decisions and behaviors during dictionary look-up can enhance successful dictionary use for language learning.

The advent of EDs has revolutionized the landscape of language learning tools, offering learners a plethora of advanced functionalities to enhance their language learning. These functionalities include hyperlinks and cross-references between related words, which enable learners to explore interconnected vocabulary and deepen their understanding of semantic relationships. Moreover, the integration of multimedia features, such as images, videos, and interactive exercises, provides learners with a more engaging and immersive learning experience, catering to diverse learning styles and preferences. In order to fully capitalize on the capabilities of EDs, learners must develop new strategies to navigate the complex array of features effectively. This necessitates a shift from traditional approaches to ED use. Additionally, learners can personalize their settings to align with their individual learning preferences, optimizing the utility of the ED for their specific needs. The portability of EDs, with their compatibility on smartphones and computers, further enhances their accessibility and convenience for learners. Instant access to look-ups enables learners to streamline their language learning process, saving valuable time compared to manually flipping through pages. Accordingly, the utilization of EDs involves a multifaceted process, demanding proficient users to possess a repertoire of adaptable skills to streamline their searches efficiently (Elola et al. 2008; Fraser 1999; Gavriilidou 2013; Scholfield 1999). Successful dictionary use has been correlated with users' capacity to develop strategies aimed at enhancing the speed of their consultations and expanding the breadth of information obtained during searches (Gavriilidou 2014). Thus, examining the strategies learners employ during ED searches can offer valuable insights for both learners and educators, aiming to promote the adoption of strategies that enhance proficient dictionary use (Mavrommatidou et al. 2019).

Drawing on Chadjipapa et al. (2020: 447) that "dictionary use is a complex process which requires the development of particular types and combinations of dictionary use strategies in different learning and cultural situations", it is reasonable to posit that learners likewise require specific types and combinations of ED use strategies to effectively utilize EDs across various learning and cultural contexts. Due to the absence of reliable methods for evaluating users' skills, characteristics, and strategies in selecting and using EDs, Mavrommatidou et al. (2019) were motivated to create a questionnaire to investigate ED use strategies. This effort led to the development and validation of the S.I.E.D.U, specifically crafted to gauge the strategies of Greek users in conducting ED searches. However, the applicability of the S.I.E.D.U in other cultural contexts remains unknown.

Examining Chinese learners' ED use strategies holds significant implications, as it enables researchers to conduct comparative analyses across cultural

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1914 (Article)

groups, elucidating potential reasons for discrepancies in ED usage. Understanding these strategies can assist educators in comprehending the constructs of ED use strategies, facilitating the creation of tailored teaching materials to address deficiencies in ED usage, thereby enhancing learners' awareness and proficiency in effective EDs utilization. For learners, comprehending their strategic EDs use allows them to assess their current skill set and identify areas for improvement, thus enhancing their ED use strategies over time. Due to significant cultural differences between China and Greece, it is expected that Chinese learners may exhibit distinct patterns in their ED usage. Moreover, the Chinese government's emphasis on digital literacy development in education since 2020 underscores the significance of EDs as learning tools. This commitment reflects the broader vision of cultivating a learning society and nation, advocating for lifelong learning opportunities. Though it does not necessarily mean that Chinese learners are already more accustomed to EDs as a digital resource, it highlights the relevance and timeliness of studying ED use strategies within the Chinese educational context, as there is an ongoing need to integrate these resources effectively into learners' habits and practices.

3. Methods

Participants were initially selected. Following this, the S.I.E.D.U. was introduced, and a translation procedure was conducted. Adaptations were made to ensure comprehensive understanding of the Chinese version of the S.I.E.D.U. Once the final version of the questionnaire was achieved, it was administered to participants.

3.1 Participants

Five hundred and eighteen English majors in their third year (seniors) from a university in South China participated in this study. All participants were native speakers of Mandarin Chinese, aged between 21 and 22. English majors were intentionally chosen due to their heavy reliance on English dictionaries throughout their English learning process. As seniors, they have been using electronic dictionaries for several years and have developed specific strategies. Their strategies in ED use can offer insights into the broader context in China. Excluding invalid questionnaire answers, 494 valid responses were collected, with 102 males and 392 females among the participants.

A larger sample size improves the accuracy and stability of factor loadings, ensures the factor solution is robust and replicable, and reduces sampling error. It meets the common guideline of having at least 5–10 participants per survey item, which strengthens the validity of the analysis. In the present study, there are 32 items in the questionnaire, meaning at least 320 participants are necessary to ensure the robustness of the study. Furthermore, a large and diverse sample increases the generalizability of the results to the broader population, ensuring that the identified factors are not specific to a small, potentially biased sample.

3.2 Introduction to the S.I.E.D.U.

The S.I.E.D.U. (Mavrommatidou et al. 2019), a self-reported instrument, consists of 32 items questionnaire designed to investigate the strategies and practices utilized by ED users. Response options are organized on a five-point Likert scale ranging from 1 = "never", to 5 = "always". Through factor analysis, researchers grouped ED use strategies included in the instrument into four distinct categories, i.e. familiarity with different types of electronic dictionaries and the conditions of their use, strategies for lemmatization and acquaintance with dictionary conventions, navigation skills, and look-up strategies in new electronic environments.

3.3 Translation of the S.I.E.D.U.

The English version of S.I.E.D.U. was originally developed in Greek (Mavrommatidou et al. 2019). In the Chinese context, we deemed it acceptable to translate the questionnaire from English to Chinese. To ensure precise translation, a group of six specialists, who are proficient in both English and Chinese, were invited. Following the translation protocol outlined by Gavriilidou (2014) and Moreira et al. (2022), the translation process of the S.I.E.D.U. into Chinese involved several meticulous steps.

Firstly, two bilingual specialists translated the questionnaire using the "decentering" method, ensuring a nuanced interpretation rather than a literal translation of each item. Cultural adaptation was also incorporated to enhance relevance to Chinese-speaking individuals. For instance, in Item 2 "I can find the dictionary I am looking for using a search engine (e.g. Google)", the Chinese version included Baidu, one of the most popular search engines in China, alongside Google, so as to enhance contextual relevance for Chinese learners. This adaptation aimed to accurately capture the intended meaning of the questionnaire items within the cultural context of the target audience. Additionally, two specialists in lexicography clarified the meanings of the content to provide well-informed recommendations for translation.

Secondly, a back translation into English was conducted by two experts to verify accuracy. Each item of the Chinese translation was back-translated into English, and the resulting English version was compared to ensure correspondence with the original English questionnaire.

Thirdly, the translated Chinese version underwent a thorough evaluation by the six specialists. This systematic approach ensures the reliability and validity of the translated S.I.E.D.U. questionnaire for Chinese-speaking populations, maintaining fidelity to the original while accounting for linguistic and cultural nuances.

3.4 Adaptation of the S.I.E.D.U.

To refine the Chinese version of the S.I.E.D.U., a pilot test was conducted with ten English majors to identify any unclear words, phrases, or sentences in the questionnaire. Eight students encountered a common issue, i.e. uncertainty regarding certain technical terms, such as "词条" (entry), "词域" (field), "通配符" (wildcards), and "布尔逻辑检索" (Boolean search). Additionally, six students expressed uncertainty about understanding the intention of specific items, such as Item 15 "To find a word in an online dictionary, I use the menu or select the first letter of a menu list using the mouse". They were unsure about the origin of the menu list mentioned in the item. In response to the feedback from the pilot study, the group of specialists decided to enhance clarity by providing explanations for these technical terms and even including screenshots of certain examples to offer further clarification. For example, we added an explanation of "当检索一个词时，只需在检索栏输入首字母，然后下拉框中会给出很多单词选项，我会看看这些选项中是否有我要查找的单词。如下图所示" to Item 15 (see Appendix for more information). Additionally, for items that remained unclear, the specialists incorporated examples to show how certain functions work. To ensure comprehensive understanding, five additional English majors were invited to review the content of the Chinese version and assess their comprehension of all items' meanings and intentions. After thorough review and confirmation, it was necessary to intentionally disrupt the order of all the items as well. Following this adjustment, the Chinese version of the S.I.E.D.U. was finalized. This proactive approach aimed to improve the comprehensibility and reliability of the questionnaire, ensuring accurate data collection in the subsequent phases of the study.

3.5 Procedure

To maintain clarity and confidentiality, the survey commenced with a concise statement outlining the study's objectives and a guarantee of data privacy. The Chinese version of the S.I.E.D.U. was integrated into the Questionnaire Star app, a popular tool used for creating and distributing surveys and questionnaires, providing accessibility to the target participants via a questionnaire link or QR code. This online questionnaire was disseminated directly to participants with the assistance of their teachers through WeChat, a social media app popular in China. Participants completed the questionnaire in a single class session, typically within a twenty-minute timeframe. Furthermore, ten out of the 494 participants were selected at random to participate in an interview, during which they shared their comprehension of the questionnaire and their strategies for using EDs.

4. Data analysis

4.1 Factor analysis

To assess the internal structure of the Chinese version of the S.I.E.D.U., we performed an exploratory factor analysis (EFA) utilizing principal component factoring with Varimax rotation, employing IBM SPSS Statistics 26. The Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy yielded a value of .887, indicating that the data were suitable for factor analysis. Additionally, Bartlett's test of sphericity was highly significant (χ^2 = 7211.505, df = 496, p = .000), further supporting the factorability of the correlation matrix. Following common practice (Hair et al. 2010), we considered factor loadings above 0.3 as indicative of acceptable item retention. Consequently, the EFA revealed the extraction of seven factors, providing a coherent grouping of the 32 questionnaire items (see Table 1).

Table 1: Factor loadings of the Chinese version of the S.I.E.D.U.

Items	**Factor Loading**
F1_10: 我能够理解电子词典中一个词条的超链接是什么，并且通过点击这个超链接，我会得到什么样的相关信息。（如下图中鼠标处） I can understand which the hyperlinks of a lexicographical entry are and where they refer to.	.746
F1_23: 我通过使用超链接来查找词条的更多信息。 I use the hyperlinks to find more information about the lexicographical entry.	.677
F1_16: 在说话过程中，为了核对一个词或短语的发音，我使用电子词典中的合成语音（电子词典所提供的单词发音功能，比如百度词典发音是合成的语音）或录音发音应用。（如下图所示） To check the pronunciation of a word/phrase while speaking, I use the application of synthesized speech or recorded pronunciation of my electronic dictionary.	.652
F1_2: 我可以通过使用搜索引擎（如谷歌，百度）找到我需要的电子词典。 I can find the dictionary I am looking for using a search engine (e.g. Google, Baidu).	.646
F1_20: 我能够在电子词典的不同功能之间轻松浏览，检索相关信息（例如单词发音、单词的释义、单词的例句、近义词、反义词等，可在这些功能之间轻松查找信息）。 I navigate/browse easily between different parts of lexicographical data.	.628
F1_32: 我使用″历史记录″选项来查看我最近进行的搜索。（如下图所示） I use the option "History" to have access to the most recent searches I carried out.	.547

F1_13: 为了更快速地检索一个单词，我在在线词典的搜索框中输入关键词，这些关键词和我要查找的词相关。 In order to search quickly for a word, I write down (in the search box of my online dictionary) keywords which are more relevant to the data of my research.	.540
F1_18: 当听到一个我不理解的单词时，我会利用电子词典中的 "Did-you-mean?" 功能进行查找，即使我不知道它的正确拼写。（例如我想搜索 nostalgia，但是在输入时输成了 nastalgia，此时搜索自动提示会问你是不是要搜索 nostalgia，即正确的词。如下图所示） When listening to a word I do not understand, I look it up even without knowing the proper spelling, utilizing the "Did-you-mean?" function of my electronic dictionary.	.524
F1_6: 我可以通过键入特定的网址来找到我需要的电子词典。 I can find the dictionary I am looking for by typing a specific URL.	.513
F2_25: 为了在在线词典中查找一个单词，我会尝试声音检索（即对着电子词典读出该单词，由电子词典自行检索该词）。 To find a word in an online dictionary, I attempt sound search.	.688
F2_4: 为了在在线词典中查找一个单词，我更喜欢使用布尔逻辑检索（即通过使用 AND, OR, NOT 等词）。 To search for a word in an online dictionary, I prefer a Boolean search (using AND, OR, NOT).	.667
F2_15: 为了在在线词典中查找一个单词，我会使用菜单或用鼠标选择菜单列表的首字母。（当检索一个词时，只需在检索栏输入首字母，然后下拉框中会给出很多单词选项，我会看看这些选项中是否有我要查找的单词。如下图所示） To find a word in an online dictionary, I use the menu or select the first letter of a menu list using the mouse.	.654
F2_22: 为了在在线词典中查找一个单词，我会使用通配符【例如问号（?），点（.），星号（*），加号（+），百分号（%）】。（例如查找like一词，我记不清词中间的字母是什么，我可以输入1??e来进行检索。如下图所示） To search for a word in an online dictionary, I use wildcards [e.g. question mark (?), dot (.), asterisk (*), plus (+), percent (%)].	.637
F2_7: 为了在在线词典中查找词组，我会尝试通过筛选进行搜索，例如按词性、词域（词属于哪个领域）、使用频率等。 To find groups of words in an online dictionary, I attempt filtered search e.g. by part of speech, field, frequency of use, etc.	.596
F2_29: 为了在在线词典中查找一个单词，我会尝试检索它的派生形式（如：happy, happily, happiness）。 To find a word in an online dictionary, I attempt inflected form search.	.515
F3_12: 在使用我的新电子词典之前，我会学习介绍词典和词条（由词目及其释义等构成的整体，是词典的基本查检单位）结构的信息。 Before using my new electronic dictionary, I study the information describing the structure of the dictionary and its entries.	.688

F3_17: 在使用我的电子词典之前，我会浏览网页（电子词典的这个网页）以了解其主要结构。 Before using my electronic dictionary, I browse the webpage to understand its main structure.	.675
F3_3: 在使用我的新电子词典之前，我会仔细学习缩略词列表（如果有的话）。 Before using my new electronic dictionary, I carefully study the list of abbreviations (if there are any).	.669
F3_28: 当电子词典提供的信息很少或可疑时，我会查找或使用纸质词典。 When the electronic dictionary has few or dubious information, I am looking for/I resort to a printed dictionary.	.568
F3_8: 我使用”帮助”选项来解决可能遇到的问题。 I use the option "Help" to solve questions and problems I may encounter.	.524
F4_11: 我使用DVD-ROM或CD-ROM上的电子词典。 I use an electronic dictionary in DVD-ROM or CD-ROM.	.759
F4_24: 我知道DVD-ROM或CD-ROM形式的电子词典是什么样子的。 I know what an electronic dictionary in DVD-ROM or CD-ROM form is.	.738
F4_31: 我知道如何将DVD-ROM中的电子词典安装到我的电脑上。 I know how to install an electronic dictionary in DVD-ROM into my computer.	.730
F5_14: 我在工作场所（学校、大学等）使用电子词典。 I use an electronic dictionary in my workplace (school, university etc.).	.753
F5_27: 我知道手机或平板电脑上的电子词典是什么。 I know what an electronic dictionary in a mobile phone or tablet is.	.741
F5_30: 我在家里使用电子词典。 I use an electronic dictionary at home.	.739
F5_1: 我知道什么是在线词典。 I know what an online dictionary is.	.678
F6_19:（和纸质词典相比），我使用电子词典更快地查找到我想要的信息。 I use an electronic dictionary to look for the desired information more quickly (compared to a print dictionary).	.704
F6_9:（和纸质词典相比），我使用电子词典更容易地查找我想要的信息。 I use an electronic dictionary to find more easily the information I want (compared to a print dictionary).	.679
F6_21: 我选择使用电子词典，因为它包含许多多媒体应用（音频，视频等），给人印象深刻。 I choose an electronic dictionary, because it contains many multimedia applications (video, audio, etc.) and is most impressive.	.632
F7_5: 我使用无需订阅的在线词典。 I use online dictionary available without subscription.	.826
F7_26: 我使用需要订阅的在线词典。 I use online dictionary available by subscription.	-.628

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1914 (Article)

The factors extracted from the Chinese version of the S.I.E.D.U. revealed a departure from the original questionnaire, necessitating careful scrutiny of each factor to ascertain its distinct characteristics. In sum, the identified seven factors collectively explain 60.98% of the total variance.

The first factor, comprising 26.663% of the total variance, encompasses nine items pertaining to users' familiarity with ED conventions and functions. Accordingly, it was labeled as "Familiarity with ED conventions and functions". The second factor, explaining 13.183% of the variance, encompasses six items focusing on users' proficiency in utilizing specific functions offered by EDs. This factor was designated as "Look-up strategies in EDs". The third factor, contributing 6.113% to the total variance, comprises five items related to users' preparation before utilizing an ED and their problem-solving approaches when encountering difficulties. This factor was labeled "Preparation and troubleshooting during ED use". The fourth factor, explaining 4.013% of the variance, consists of three items concerning the storage format of EDs, specifically whether it can be gained from DVD-ROM or CD-ROM. It was labeled "Storage format of EDs (DVD-ROM/CD-ROM)". The fifth factor, contributing 3.868% to the total variance, encompasses four items related to users' overall acceptance of EDs and the contexts in which they are utilized, earning the label "Acceptance and usage context of EDs". The sixth factor, explaining 3.709% of the variance, comprises three items focusing on the advantages that EDs offer over traditional paper dictionaries. This factor was designated "Advantages of ED over traditional methods". Lastly, the seventh factor, accounting for 3.431% of the variance, consists of two items associated with users' subscriptions to ED services. It was labeled "Subscription and access to EDs".

By meticulously delineating each factor and its constituent items, this analysis provides a nuanced understanding of the underlying dimensions of ED usage among the studied population group.

4.2 Internal consistency and reliability

Reliability in assessment is crucial as it ensures the consistency and stability of questionnaire data (Johnson and Christensen 2000). Cronbach's alpha, a widely utilized measure of internal consistency reliability, evaluates how closely related a set of items are as a group within a scale. This coefficient ranges from 0 to 1, with values of 0.9 and above indicating excellent reliability, 0.8 to 0.89 signifying good reliability, 0.7 to 0.79 suggesting acceptable reliability, 0.6 to 0.69 implying questionable reliability, 0.5 to 0.59 indicating poor reliability, and values below 0.5 considered unacceptable (Cronbach 1951). To assess the internal consistency and reliability of the Chinese version of the S.I.E.D.U., reliability testing was conducted for both the overall scale and each factor (see Table 2).

It is imperative to note that the value of factor 7 is negative, stemming from a negative average covariance among items. This deviation violates the assumption of the reliability model. The reasons for this anomaly are discussed in the following section. It is recommended to remove the items included in factor 7 from the questionnaire. Upon deletion of Factor 7, a higher Cronbach's alpha value is obtained, thereby enhancing the overall reliability of the scale.

Table 2: Internal consistency reliability for the overall scale and each factor

Scales	**Cronbach's alpha**
Total scale	.902
Total scale after deleting Factor 7	.905
Factor 1: Familiarity with ED conventions and functions	.849
Factor 2: Look-up strategies in EDs	.831
Factor 3: Preparation and troubleshooting during ED use	.822
Factor 4: Storage format of EDs (DVD-ROM/CD-ROM)	.756
Factor 5: Acceptance and usage context of EDs	.786
Factor 6: Advantages of ED over traditional methods	.695
Factor 7: Subscription and access to EDs	-1.003

5. Discussion

In contrast to the four factors identified in the S.I.E.D.U., the Chinese version revealed seven distinct factors. Within these seven factors, both commonalities and differences emerged when comparing the findings of the two versions (see Table 3). It was noted that the four factors derived from the original study exhibit some degree of overlap. Specifically, factors 1 and 2 pertain to ED conventions, while factors 3 and 4 delve into learners' reference skills and strategies employed during ED consultations.

Table 3: Comparison of the factors extracted from the Chinese version of the S.I.E.D.U. and its original counterpart

Factors of the Chinese version of the S.I.E.D.U.	**Factors of the S.I.E.D.U.**
Factor 1: Familiarity with ED conventions and functions	Factor 1: Familiarity with different types of electronic dictionaries and the conditions of their use
Factor 2: Look-up strategies in EDs	Factor 2: Strategies for lemmatization and acquaintance with dictionary conventions
Factor 3: Preparation and troubleshooting during ED use	Factor 3: Navigation skills
Factor 4: Storage format of EDs (DVD-ROM/CD-ROM)	Factor 4: Look-up strategies in new electronic environments

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1914 (Article)

Factor 5: Acceptance and usage context of EDs
Factor 6: Advantages of ED over traditional methods
Factor 7: Subscription and access to EDs

In our survey, factors 1 and 2 exhibit congruent meanings with those identified in the prior study, highlighting the critical importance of familiarity with ED conventions and adept look-up strategies. These two factors are foundational pillars that underpin effective ED searches. Familiarity with ED conventions involves grasping their structure, functionalities, and features, as well as how entries are organized, abbreviations utilized, available search options, and more. Mastery of these conventions empowers users to navigate ED interfaces with ease, maximizing efficiency and accuracy in retrieving desired information. Similarly, proficiency in look-up strategies involves employing systematic approaches to identify and locate information within an ED efficiently. This includes selecting appropriate search terms, utilizing advanced search features, and interpreting search results effectively. Together, factors 1 and 2 constitute a substantial portion of the total variance, amounting to 39.846%. This underscores the pivotal role that familiarity with ED conventions and adept look-up strategies play in facilitating successful ED searches. As foundational competencies, they equip learners with the essential skills and knowledge needed to harness the full potential of EDs, enhancing their effectiveness and utility in various academic and practical contexts.

However, the present study unveils more nuanced and intriguing findings. Factor 3 emerged to capture learners' behaviors prior to utilizing a new ED and their approaches to troubleshooting when encountering difficulties. Interestingly, preparations before ED usage are often overlooked by learners. As mentioned by Interviewee 4, "I seldom do anything before using a new electronic dictionary; I thought dictionaries are generally the same". It is acknowledged that a grasp of the ED's structure, entry display, abbreviations, and other features can expedite the consultation process and facilitate obtaining desired information for the learners themselves. The challenges encountered during searches can partly be attributed to learners' lack of fundamental knowledge about EDs. This result is consistent with previous research, which calls for dictionary use training. Despite not being as prominent as the preceding factors, Factor 3 still contributes significantly to the overall variance.

Factor 4 in the current study addresses the storage format of electronic dictionaries, specifically in DVD-ROM or CD-ROM formats. Notably, during the interviews, eight out of the ten participants expressed views on the declining relevance of DVD-ROMs and CD-ROMs in today's context. A significant observation was the absence of drivers for these formats on their computers or laptops, indicating a lack of usage or even possession of such physical media among many interviewees. Some even viewed DVD-ROMs and CD-ROMs as cumbersome and outdated. This trend is multifaceted. Firstly, the limited accessibility of physical media-based dictionaries poses significant inconvenience, requiring specific devices like computers or DVD/CD drives for access. In an age where learners prioritize convenience and portability, the need for specialized hardware can deter usage. Additionally, technological advancements have rendered DVD-ROMs and CD-ROMs obsolete, with modern learners favoring web-based platforms and mobile apps for their flexibility and compatibility across devices. Moreover, physical media-based dictionaries lack the interactive features and real-time updates offered by digital alternatives, failing to meet the expectations of today's tech-savvy learners. Cost considerations also play a role, as physical media entail upfront expenses and additional shipping costs, making them less attractive compared to online alternatives. In brief, the declining popularity of DVD-ROM or CD-ROM electronic dictionaries can be attributed to a combination of technological advancements, changing user preferences, and the availability of more convenient digital options.

Factor 5 explores the acceptance and usage contexts of EDs, revealing a widespread embrace of these tools by learners. This factor suggests a prevailing trend wherein learners increasingly rely on EDs to fulfill their linguistic needs. For instance, Interviewee 9's statement, "I don't have a paper dictionary in my college life, all I use is Internet and electronic dictionaries", vividly illustrates this preference for digital resources. Similarly, Interviewee 5's remark, "As an English major, I have to use dictionaries frequently, but all I use is an electronic dictionary because it's faster and more convenient", echoes this sentiment, highlighting the collective shift towards the convenience and accessibility offered by EDs. This trend aligns with Sharpe's (1995: 49) observation that "the familiarity of today's young people with electronic devices will eventually relegate the printed notion of 'dictionary' to a secondary sense". Furthermore, the adaptability of EDs is evident through their versatility across various contexts, transcending the boundaries of academia and seamlessly integrating into learners' daily lives. Whether utilized in the workplace, school, university, or at home as mentioned in Items 14 and 30, for purposes ranging from language study to everyday communication, EDs have become indispensable tools, readily accessible whenever and wherever needed. This pervasive integration underscores the transformative impact of EDs, reshaping the landscape of language learning and usage in contemporary society.

Factor 6 comprises three out of four items containing the phrase "compared to a printed dictionary", while the last item, although lacking this comparison, also outlines the reasons why learners prefer EDs. For example, Item 19 highlights that learners can search for desired information more quickly, Item 9 emphasizes the ease of finding information, and Item 21 underscores the presence of multimedia applications in EDs. Overall, these four items effectively underline the advantages of EDs over their printed counterparts, showcasing their superiority in terms of speed, ease of use, and multimedia functionality. Alamri and Hakami's (2022) research supports this, showing that EFL learners prefer EDs over printed dictionaries due to quicker information access, time efficiency, and accurate language translation. Through this factor, valuable insights into the reasons behind learners' acceptance of EDs and their widespread popularity across various usage contexts can be gleaned.

Factor 7 comprises two items related to ED subscriptions, with factor loadings for Item 5 and Item 26 indicated in Table 1. It is noteworthy that Item 26 presents the opposite statement to Item 5, which explains why the factor loading of Item 26 yields a negative value. Given that both items essentially explore the same aspect, the presence of positive and negative statements inevitably contributes to result inconsistency, necessitating their removal. Nevertheless, this observation highlights a significant trend that participants generally prefer not to subscribe to EDs. In the dynamic landscape of contemporary learning environments, learners are increasingly eschewing ED subscriptions for various reasons. Firstly, the proliferation of free alternatives, including online dictionaries and translation tools, provides learners with readily accessible resources without the encumbrance of subscription fees. Furthermore, lingering doubts regarding the quality and comprehensiveness of subscription-based EDs compared to their free counterparts may weigh heavily on learners' decision-making processes. Economic considerations also wield significant influence, with learners conscientiously managing their educational expenditures and favoring free resources over subscription-based options. Moreover, the ever-expanding array of information sources beyond traditional dictionaries, such as language learning apps and educational websites, serves to dilute the perceived indispensability of subscription-based EDs. This shifting landscape is further propelled by technological preferences, with learners gravitating towards mobile apps and web-based platforms for their unparalleled convenience and ubiquitous accessibility. Consequently, the perceived value proposition of subscription-based EDs may fail to resonate with the diverse needs and preferences of modern learners, compelling them to explore and adopt alternative language tools and resources that better align with their evolving learning paradigms. In light of this perspective, exploring learners' views on subscribing to EDs becomes meaningless. Therefore, it is recommended to remove Item 5 and Item 26 from the questionnaire.

Compared to the four factors identified in the S.I.E.D.U., the seven factors revealed in the present study offer more nuanced and insightful findings. In addition to factors correlated to ED conventions, functions, and strategic skills, our study identified additional factors including learners' preparation and troubleshooting, acceptance and usage context, storage format and advantages of EDs, and ED subscription. Together, these factors provide a comprehensive understanding of ED use strategies from the Chinese learners' perspective, benefiting not only learners but also educators.

6. Conclusion

The motivation behind adapting and validating the S.I.E.D.U. was to assess its

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1914 (Article)

applicability not only in the Greek context but also in the Chinese setting. As part of this process, two items related to the subscription of EDs are recommended for removal based on our factor analysis results. The decision to exclude these items in the Chinese context reflects the diverse needs and preferences of modern learners. Despite this adjustment, the majority of the items confirmed the S.I.E.D.U. as a reliable and valid instrument for evaluating the ED use strategies in Chinese culture. The present study yielded more nuanced results, offering insightful views on Chinese learners' ED use strategies. The effective utilization of EDs requires users to develop proficiency not only in navigating the interface but also in discerning relevant information and integrating it into their language learning endeavors. Accordingly, these findings provide valuable pedagogical implications that can inform educational practices and interventions in the realm of language learning. Recommendations for using the Chinese version in China include adopting it to diverse educational settings to optimize its effectiveness. Moreover, adapting the original version for use in other linguistic and cultural contexts would require additional validation and modifications to ensure the accuracy and applicability of the S.I.E.D.U. across different cultural backgrounds. Future research endeavors could further investigate its adaptability across various linguistic and cultural settings.

To optimize strategies in ED usage, Chinese learners and educators must familiarize themselves with ED conventions. During the adaptation process of the questionnaire, it was noted that learners were unfamiliar with some technical terms and certain functions of EDs. This lack of familiarity resulted in uncertainty regarding the intended meaning of specific items in the translated questionnaire. This unfamiliarity reflects Chinese learners' limited knowledge about ED conventions, which potentially impedes their effective utilization of EDs. To address this issue, Chinese learners must dedicate time to understanding the conventions and functionalities of EDs, including search options, entry organization, abbreviations, and more. By doing so, learners can enhance their efficiency in utilizing these digital resources, thereby optimizing their overall experience with EDs. Additionally, learners can develop effective look-up strategies by practicing systematic approaches, including selecting appropriate search terms and utilizing advanced search features. This will streamline the search process and improve the accuracy of information retrieval. Furthermore, learners can explore various ED platforms and leverage their multimedia functionalities for a richer learning experience. Lastly, learners should recognize the versatility of EDs and employ them across various contexts, integrating them into their daily routines beyond academic research or language study. This broader usage will allow learners to maximize the utility and effectiveness of EDs across diverse learning endeavors.

To support learners' proficiency in utilizing EDs, educators can integrate ED training into language learning curricula. Hadebe (2004) emphasizes the importance of teacher training in dictionary skills as the fundamental prerequisite to provide learners with personalized instruction tailored to their individ-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1914 (Article)

ual needs. In a similar vein, Bogaards (2003) stresses the critical importance of dictionary use training, highlighting the significant gap between the lack of progress in training programs and the advancement in dictionary quality. Gavriilidou et al. (2024) endorse the teachability of dictionary use strategies and skills, proposing that well-designed dictionary awareness programs can enhance awareness and foster a culture of dictionary use. These observations highlight the imperative for educators to include ED training in their curricula. Additionally, educators can promote critical thinking skills by encouraging learners to evaluate the quality and reliability of information retrieved from EDs. Improving the existing textbooks to better meet learners' needs for dictionary skills training is also essential (Law 2024). By teaching learners to discern credible sources, verify information, and cross-reference multiple resources, educators can enhance learners' ED use strategies. Furthermore, providing access to a diverse range of ED platforms exposes learners to different functionalities and features. Familiarizing learners with web-based, mobile app, and software-based EDs accommodates diverse learning preferences and needs. Last but not least, educators can foster collaborative learning environments where learners share tips, strategies, and resources related to ED usage. Encouraging peer-to-peer support and interaction can also promote active engagement and knowledge exchange among learners. Through these pedagogical approaches, both learners and educators can harness the full potential of EDs to enhance language learning and teaching outcomes.

Undoubtedly, this study has limitations inherent to the nature of the research tool. The Chinese version of the S.I.E.D.U. relies on self-reported responses, raising questions about the extent to which participants' answers truly reflect their real and objective perceptions (Chamot 2004). Participants might provide answers that they perceive as socially desirable or that align with their expectations of the study's goals, rather than their genuine experiences or behaviors. Future studies could consider combining self-reported data with observational or qualitative methods to explore learners' ED use strategies. Such instruments could provide a clearer and more accurate understanding of learners' behaviors and perceptions in this domain, thereby enhancing the credibility and rigor of research findings.

Acknowledgements

We are grateful to Prof. Elsabé Taljard, and two anonymous reviewers for their insightful comments and suggestions. Remaining errors are our own. Correspondence should be addressed to Hui Wang.

This study was supported by Guangdong Philosophy and Social Sciences Foundation Project (Grant No. GD23WZXC01-02), 2023 Higher Education Teaching Reform Project of Guangdong Province (Entitled "Innovative Practice of Ideological and Political Education in International Literacy English Courses Based on Output-Oriented Approach", and "Reform and Practice of AI-powered

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1914 (Article)

Teaching Model for College English in the Context of Excellent Engineers"), and Dongguan University of Technology's Teaching Reform Project of Higher Education (Grant No. 202202025).

References

Alamri, H.R. and H.M. Hakami. 2022. Exploring Perspectives of EFL Students on Using Electronic Dictionaries to Improve Vocabulary Learning: A Comparative Study. *International Journal of Curriculum and Instruction* 14(2): 1578-1599.

Bogaards, P. 2003. Uses and Users of Dictionaries. Van Sterkenburg, P. (Ed.). 2003. *A Practical Guide to Lexicography, Terminology and Lexicography Research and Practice* 6: 26-33. Amsterdam/Philadelphia: John Benjamins.

Campoy-Cubillo, M.C. 2021. Fostering Learners' Online Dictionary Skills through Active Dictionary Rubrics. *Lexikos* 31: 487-510.

Chadjipapa, E., Z. Gavriilidou, A. Markos and A. Mylonopoulos. 2020. The Effect of Gender and Educational Level on Dictionary Use Strategies Adopted by Upper-elementary and Lower-secondary Students Attending Greek Schools. *International Journal of Lexicography* 33(4): 443-462.

Chamot, A.U. 2004. Issues in Language Learning Strategy Research and Teaching. *Electronic Journal of Foreign Language Teaching* 1(1): 14-26.

Cohen, A.D. and R.L. Oxford. 2002. Young Learners' Language Strategy Use Survey. Cohen, A.D. and S.J. Weaver. 2006. *Styles and Strategies-based Instruction: A Teachers' Guide:* 60-63. Minneapolis: Center for Advanced Research on Language Acquisition, University of Minnesota.

Cronbach, L.J. 1951. Coefficient Alpha and the Internal Structure of Tests. *Psychometrika* 16(3): 297-334.

Elola, I., V. Rodríguez-García and K. Winfrey. 2008. Dictionary Use and Vocabulary Choices in L2 Writing. *ELIA* 8: 63-89.

Fan, M.Y. 2003. Frequency of Use, Perceived Usefulness, and Actual Usefulness of Second Language Vocabulary Strategies: A Study of Hong Kong Learners. *The Modern Language Journal* 87(2): 222-241.

Fraser, C.A. 1999. The Role of Consulting a Dictionary in Reading and Vocabulary Learning. *Canadian Journal of Applied Linguistics* 2(1–2): 73-89.

Gavriilidou, Z. 2011. Profiling Greek Adult Dictionary Users. *Studies in Greek Linguistics* 31: 166-172.

Gavriilidou, Z. 2013. Development and Validation of the *Strategy Inventory for Dictionary Use* (S.I.D.U.). *International Journal of Lexicography* 26(2): 135-153.

Gavriilidou, Z. 2014. Translation, Cultural Adaptation and Preliminary Psychometric Evaluation of the English Version of "Strategy Inventory for Dictionary Use" (S.I.D.U.). Abel, A., C. Vettori and N. Ralli (Eds.). 2014. *Proceedings of the XVI Euralex International Congress: The User in Focu*s, *EURALEX 2014, Bolzano/Bozen, Italy, July 15–19, 2014:* 225-235. Bolzano/Bozen: EURAC Research.

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.

Gouws, R.H. and D.J. Prinsloo. 2021. Lexicographic Data Boxes. Part 3: Aspects of Data Boxes in Bilingual Dictionaries and a Perspective on Current and Future Data Boxes. *Lexikos* 31(1), 402-433.

Gu, P.Y. 2003. Vocabulary Learning in a Second Language: Person, Task, Context and Strategies. *TESL-EJ* 7(2): 1-25.

Hadebe, S. 2004. Improving Dictionary Skills in Ndebele. *Lexikos* 14: 89-104.

Hair, J.F., W.C. Black, B.J. Babin and R.E. Anderson. 2010. *Multivariate Data Analysis*. Upper Saddle River, NJ: Prentice Hall.

Harvey, K. and D. Yuill. 1997. A Study of the Use of a Monolingual Pedagogical Dictionary by Learners of English Engaged in Writing. *Applied Linguistics* 18(3): 253-278.

Joffe, D. 2009. TLex: Setting New Standards for a Global, Fully-Integrated e-Lexicography Workbench and Electronic Dictionary Publishing System. *eLexicography in the 21st Century, eLEX 2009: New Challenges, New Applications, Louvain-la-Neuve, 22–24 October 2009. Book of Abstracts:* 109-111. De Louvain: UCL Press.

Johnson, B. and L. Christensen. 2000. *Educational Research: Quantitative and Qualitative Approaches*. Boston, MA: Allyn & Bacon.

Klein, J. 2008. Teaching Dictionary-using Skills for Online Dictionaries: An Attempt at a Theoretical Framework for South Africa. Bernal, E. and J. DeCesaris (Eds.). 2008. *Proceedings of the XIII Euralex International Congress, Barcelona, Spain, July 15–19, 2008*: 225-226. Barcelona: Institut Universitari de Lingüística Aplicada, Universitat Pompeu Fabra.

Krajka, J. 2015. Online Lexicological Tools in ESP–Towards an Approach to Strategy Training. *Scripta Manent* 3(1): 3-19.

Law, W.-O. 2024. Dictionary Use Training in Secondary School EFL Textbooks in Taiwan. *Lexikos* 34(1): 77-98.

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.

Mavrommatidou, S., Z. Gavriilidou and A. Markos. 2019. Development and Validation of the Strategy Inventory for Electronic Dictionary Use (S.I.E.D.U.). *International Journal of Lexicography* 32(4): 393-410.

McAlpine, J. and J. Myles. 2003. Capturing Phraseology in an Online Dictionary for Advanced Users of English as a Second Language: A Response to User Needs. *System* 31(1): 71-84.

Moreira, T., C. Faiad, A.D. Santos Araújo Jesuíno, A. Lima-Costa and A.C. Frenzel. 2022. Adaptation of the Teacher Emotions Scales (TES) to the Brazilian Context. *Psico-USF* 26: 71-81.

Nation, I.S.P. 2001. *Learning Vocabulary in Another Language*. Cambridge, UK: Cambridge University Press.

Nesi, H. 2000. Electronic Dictionaries in Second Language Vocabulary Comprehension and Acquisition: The State of the Art. 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. Vol. 2*: 839-847. Stuttgart: Institut für Maschinelle Sprachverarbeitung, Universität Stuttgart.

O'Malley, J.M. and A.U. Chamot. 1990. *Learning Strategies in Second Language Acquisition*. Cambridge, UK: Cambridge University Press.

Oxford, R. 1990. *Language Learning Strategies: What Every Teacher Should Know.* Rowley, MA: Newbury House.

Pastor, V. and A. Alcina. 2010. Search Techniques in Electronic Dictionaries: A Classification for Translators. *International Journal of Lexicography* 23(3): 307-354.

Rundell, M. 2013. Redefining the Dictionary: From Print to Digital. *Kernerman Dictionary News* 24: 5-7.

Scholfield, P. 1999. Dictionary Use in Reception. *International Journal of Lexicography* 12(1): 13-34.

Sharpe, P. 1995. Electronic Dictionaries with Particular Reference to the Design of an Electronic Bilingual Dictionary for English-Speaking Learners of Japanese. *International Journal of Lexicography* 8(1): 39-54.

Summers, D. 2013. The Role of Dictionaries in Language Learning. Carter, R. and M. McCarthy (Eds.). 2013. *Vocabulary and Language Teaching:* 111-125. London: Routledge.

Verlinde, S., P. Leroyer and J. Binon. 2009. Search and You Will Find. From Stand-alone Lexicographic Tools to User Driven Task and Problem-oriented Multifunctional Leximats. *International Journal of Lexicography* 23(1): 1-17.

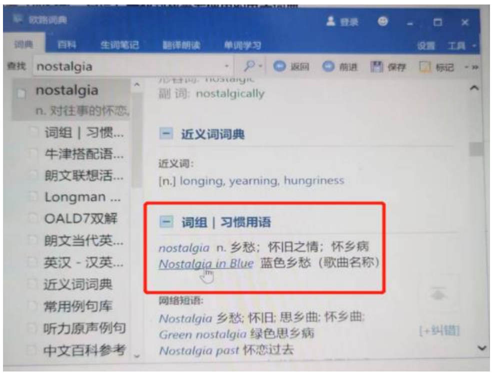
Winkler, B. 2001. *Students Working with an English Learners' Dictionary on CD-ROM. Papers from the ITMELT 2001 Conference.* Available at:
http://www2.elc.polyu.edu.hk/conference/papers2001/winkler.htm

Appendix: The Chinese version of the adapted S.I.E.D.U.

本问卷旨在调查中国英语学习者的电子词典使用情况，内容包含两部分：第一部分为个人信息，请按照提示要求如实用汉语或数字填写；第二部分为 32 项情况描述，每个描述项下方有5个选项（①不符合；②基本不符合；③一般；④大多数情况下符合；⑤ 完全符合），请依据自身情况，选择符合自身情况的描述项。请填写人认真如实填写问卷内容，您的真实数据对我们的研究结果非常重要。我们承诺，本问卷所搜集数据将严格保密，仅用于科学研究需要，在研究结果分析与汇报中也将充分保护填写人的个人隐私。

姓名
性别
年龄
学校
专业

1. 我能够理解电子词典中一个词条的超链接是什么，并且通过点击这个超链接，我会得到什么样的相关信息。（如下图中鼠标处）



nostalgia

nostalgia
n. 对往事的怀恋,

副 词: nostalgically

近义词词典

近义词:
[n.] longing, yearning, hungriness

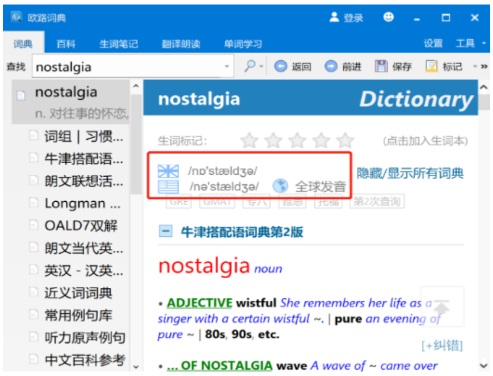
词组 | 习惯用语

nostalgia n. 乡愁；怀旧之情；怀乡病
Nostalgia in Blue 蓝色乡愁（歌曲名称）

网络短语:
Nostalgia 乡愁; 怀旧; 思乡曲; 怀乡曲;
Green nostalgia 绿色思乡病
Nostalgia past 怀恋过去

2. 我通过使用超链接来查找词条的更多信息。

3. 在说话过程中，为了核对一个词或短语的发音，我使用电子词典中的合成语音（电子词典所提供的单词发音功能，比如百度词典发音是合成的语音）或录音发音应用。（如下图所示）



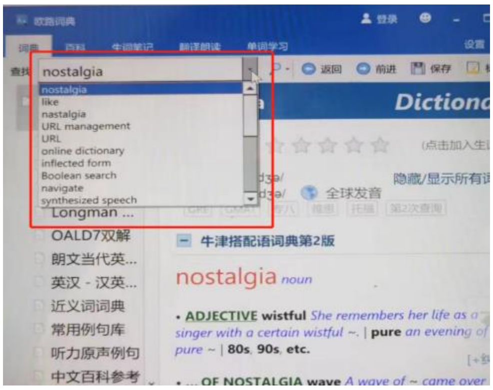
• **ADJECTIVE** wistful *She remembers her life as a singer with a certain wistful ~.* | **pure** *an evening of pure ~* | **80s, 90s, etc.**

• **... OF NOSTALGIA** wave *A wave of ~ came over*

4. 我可以通过使用搜索引擎（如谷歌，百度）找到我需要的电子词典。

5. 我能够在电子词典的不同功能之间轻松浏览，检索相关信息（例如单词发音，单词的释义，单词的例句，近义词，反义词等，可在这些功能之间轻松查找信息）。

6. 我使用”历史记录”选项来查看我最近进行的搜索。（如下图所示）



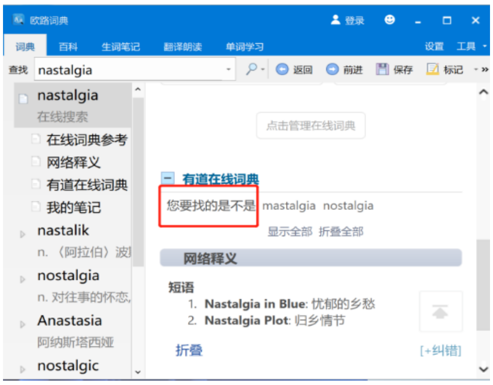
• **ADJECTIVE** wistful *She remembers her life as a singer with a certain wistful ~.* | **pure** *an evening of pure ~* | **80s, 90s, etc.**

• **... OF NOSTALGIA** wave *A wave of ~ came over*

7. 为了更快速地检索一个单词，我在在线词典的搜索框中输入关键词，这些关键词和我要查找的词相关。

8. 当听到一个我不理解的单词时，我会利用电子词典中的 "Did-you-mean?" 功能进行查找，即使我不知道它的正确拼写。（例如我想搜索 nostalgia，但是在输入时输成了 nastalgia，此时搜索自动提示会问你是不是要搜索 nostalgia，即正确的词。如下图所示）

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1914 (Article)



查找 nastalgia

nastalgia
在线搜索
在线词典参考
网络释义
有道在线词典
我的笔记
nastalik
n. （阿拉伯）波
nostalgia
n. 对往事的怀恋,
Anastasia
阿纳斯塔西娅
nostalgic

有道在线词典

您要找的是不是 mastalgia nostalgia

网络释义

短语

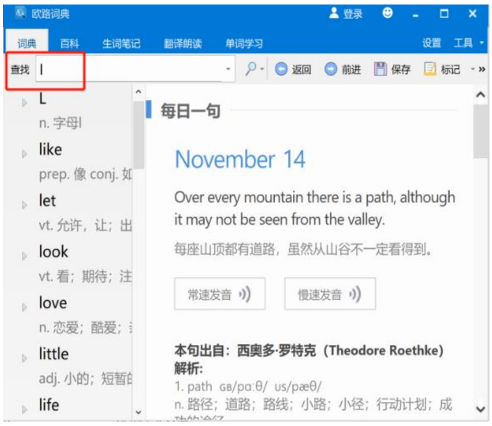
1. **Nastalgia in Blue**: 忧郁的乡愁
2. **Nastalgia Plot**: 归乡情节

9. 我可以通过键入特定的网址来找到我需要的电子词典。

10. 为了在在线词典中查找一个单词，我会尝试声音检索（即对着电子词典读出该单词，由电子词典自行检索该词）。

11. 为了在在线词典中查找一个单词，我更喜欢使用布尔逻辑检索（即通过使用 AND, OR, NOT 等词）。

12. 为了在在线词典中查找一个单词，我会使用菜单或用鼠标选择菜单列表的首字母。（当检索一个词时，只需在检索栏输入首字母，然后下拉框中会给出很多单词选项，我会看看这些选项中是否有我要查找的单词。如下图所示）



查找 l

L
n. 字母l
like
prep. 像 conj. 如
let
vt. 允许，让；出
look
vt. 看；期待；注
love
n. 恋爱；酷爱；
little
adj. 小的；短暂的
life

每日一句

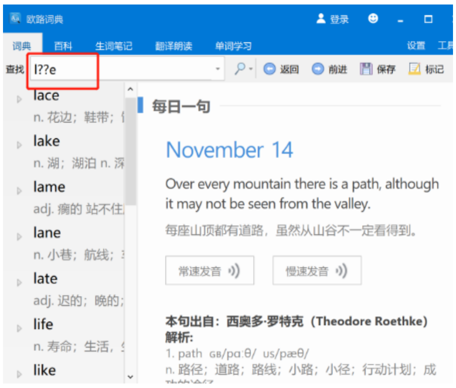
November 14

Over every mountain there is a path, although it may not be seen from the valley.

每座山顶都有道路，虽然从山谷不一定看得到。

本句出自：西奥多·罗特克（Theodore Roethke）
解析:
1. path GB/pɑːθ/ US/pæθ/
n. 路径；道路；路线；小路；小径；行动计划；成

13. 为了在在线词典中查找一个单词，我会使用通配符【例如问号（?），点（.），星号（*），加号（+），百分号（%）】。（例如查找like一词，我记不清词中间的字母是什么，我可以输入l??e来进行检索。如下图所示）



l??e

lace
n. 花边；鞋带；

lake
n. 湖；湖泊 n. 深

lame
adj. 瘸的 站不住

lane
n. 小巷；航线；

late
adj. 迟的；晚的；

life
n. 寿命；生活，

like

每日一句

November 14

Over every mountain there is a path, although it may not be seen from the valley.

每座山顶都有道路，虽然从山谷不一定看得到。

本句出自：西奥多·罗特克（Theodore Roethke）
解析:
1. path GB/pɑːθ/ US/pæθ/
n. 路径；道路；路线；小路；小径；行动计划；成

14. 为了在在线词典中查找词组，我会尝试通过筛选进行搜索，例如按词性、词域（词属于哪个领域）、使用频率等。

15. 为了在在线词典中查找一个单词，我会尝试检索它的派生形式（如：happy, happily, happiness）。

16. 在使用我的新电子词典之前，我会学习介绍词典和词条（由词目及其释义等构成的整体，是词典的基本查检单位）结构的信息。

17. 在使用我的电子词典之前，我会浏览网页（电子词典的这个网页）以了解其主要结构。

18. 在使用我的新电子词典之前，我会仔细学习缩略词列表（如果有的话）。

19. 当电子词典提供的信息很少或可疑时，我会查找或使用纸质词典。

20. 我使用”帮助”选项来解决可能遇到的问题。

21. 我使用DVD-ROM或CD-ROM上的电子词典。

22. 我知道DVD-ROM或CD-ROM形式的电子词典是什么样子的。

23. 我知道如何将DVD-ROM中的电子词典安装到我的电脑上。

24. 我在工作场所（学校、大学等）使用电子词典。

25. 我知道手机或平板电脑上的电子词典是什么。

26. 我在家里使用电子词典。

27. 我知道什么是在线词典。

28.（和纸质词典相比），我使用电子词典更快地查找到我想要的信息。

29.（和纸质词典相比），我使用电子词典更容易地查找我想要的信息。

30. 我选择使用电子词典，因为它包含许多多媒体应用（音频、视频等），给人印象深刻。

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1926 (Article)

Using Semi-automated Term Extraction for IsiNdebele Health Terminology

Nomsebenzi Malele, *Department of African Languages,*
University of South Africa, Pretoria, South Africa
(malelnj@unisa.ac.za) (https://orcid.org/0000-0001-8384-7853)
and
Sonja Bosch, *Department of African Languages,*
University of South Africa, Pretoria, South Africa
(seb@hbosch.com) (https://orcid.org/0000-0002-9800-5971)

Abstract: IsiNdebele, also known as Southern isiNdebele, has a limited availability of language resources and specialised terminology, especially when compared to other members of the Nguni language family. This study therefore explores means of addressing the shortage of specialised terminology in isiNdebele by using semi-automatic term extraction methods. The focus is on health terminology, intended for communication with laypersons rather than between experts in the health field. Semi-automatic term extraction methods are employed, combining manual identification and extraction of data from available corpora with the use of a software tool named WordSmith Tools (WST). The study illustrates the necessity of utilising all functions of the WST, as they complement each other. Terms overlooked by one function may be captured by another. For instance, while the KeyWords function identified only a limited number of terms in this research, manual identification proved more fruitful. Interestingly, the Concord function emerged as particularly effective in identifying a greater number of terms. The use of the WST in this research highlights the viability of corpus-driven studies, even for resource-scarce languages like isiNdebele. Therefore, considering the limited resources available for isiNdebele, particularly the absence of specialised dictionaries, this collection of health terms exemplifies ideal candidates for inclusion in a general dictionary.

Keywords: ISINDEBELE, CORPUS-DRIVEN TERM EXTRACTION, HEALTH CORPORA, LANGUAGE FOR SPECIFIC PURPOSES (LSP), LANGUAGE FOR GENERAL PURPOSES (LGP), WORDSMITH TOOLS, WORD LIST, KEY WORDS, CONCORDANCE, SEMI-AUTOMATIC EXTRACTION

Irhunyezorhubhululo: Ukusebenzisa Indlela Yemitjhini Nezandla Ukutsomula Itheminoloji yesiNdebele Yezamaphilo. IsiNdebele, esibuye saziwe ngokobana siNdebele seSewula, sitlhayelelwa khulu mithombo yelimi kanye netheminoloji ekhethekileko khulukhulu, lokha umuntu nakasimadanisa namanye amalimi wabeNguni, isiNdebele esiyingcenye yawo. Ngalokho irhubhululweli lihlola iindlela zokuhlangabezana nalokhu kutlhayela kwetheminoloji ekhethekileko esiNdebeleni. Lokhu kwenziwa ngokusebenzisa iindlela zokutsomula amathemu kusetjeniswa imitjhini nezandla. Umnqopho werhubhululweli usetheminolojini yezamaphilo. Kuhloswe bona imikhulumiswano ibe lula hlangana nabantu abangasi lilitho kunokobana kube nemi-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1926 (Article)

khulumiswano elula hlangana nabocwephetjhe bomkhakha wezamaphilo. Njengombana sekuveziwe ngehla ukobana kusetjenziswa indlela yokutsomula amathemu ngomtjhini nangezandla, kilelirhubhululo, umtjhini osetjenzisweko ubizwa bona yiWordSmith Tools. Lelirhubhululo litjengisa ukuqakatheka kokusetjenziswa kwawo woke amathulusi weWordSmith Tools (WST) ngombana, womathathu aphelelisana kuhle khulu. Lokho kutjho bona amathemu angakhange alemukwe ngelinye ithulusi, ayalemukwa ngelinye. Isibonelo, njengombana iKeyWords ikghone ukulemuka amathemu ambalwa kangaka, ithulusi iConcord lona likwazile ukulemuka amathemu amanengi ngendlela erarako. Ukusetjenziswa kwe-WST kilelirhubhululo kuveza ngokusobala ukusebenza kuhle kwerhubhululo elisunduzwa yikhophasi nemalimini atlhayelelwa ziinstjenziswa njengesiNdebele. Ngalokho lokha nawutjhejisisa ukutlhayelelwa kwelimeli zizinto ezifana neenhlathululimezwi ezikhethekileko, lokhu kubuthelelwa kwamathemu wezamaphilo kuveza lawo mathemu angahle afakwe kusihlathululimezwi esivamileko.

Amagama aqakathekileko: ISINDEBELE, UKUTSOMULWA KWAMATHEMU OKUSUNDUZWA YIKHOPHASI, IKHOPHORA YEZAMAPHILO, ILIMI LOMNQOPHO OKHETHEKILEKO (LSP), ILIMI LOMNQOPHO OVAMILEKO (LGP), IWORDSMITH TOOLS, IWORD LIST, IKEYWORDS, ICONCORDANCE, UKUTSOMULA NGOMTJHINI NANGEZANDLA

1. Introduction and background

IsiNdebele, also known as Southern Ndebele (ISO 639-3: ndl)[1], is one of the twelve official languages of South Africa and is primarily spoken in the former kwaNdebele region of Mpumalanga. It belongs to the Nguni group of languages which includes isiZulu (zul), isiXhosa (xho), Zimbabwean Ndebele (nde) and Siswati (ssw). IsiNdebele exhibits a morphological complexity based on a robust noun class system resulting in the extensive use of prefixes and suffixes. These morphological elements play a crucial role in shaping the meaning of words and sentences. In Nguni languages a conjunctive orthography is employed which is characterised by a seamless and interconnected representation of linguistic elements (Taljard and Bosch 2006: 432-433), whereas the Sotho group of languages, including Sesotho sa Leboa (also known as Northern Sotho or Sepedi) employs a disjunctive orthography that introduces distinct boundaries between linguistic units.

Despite being a long-established spoken language, isiNdebele was first given full written status in 1985, when it was first introduced in classrooms. Prior to 1985, isiZulu rather than isiNdebele was the language of teaching for the children of the Ndebele people. In 1996, the language was examined for the first time as a subject for matriculation (Jiyane 1994: 1). In comparison to other official languages, especially those of the Nguni language family, isiNdebele performs poorly in terms of language resources. There is no language for specific purposes (LSP) dictionary available for isiNdebele. Only general-purpose (LGP) dictionaries exist, indicating the scarcity of specialised terminology in isiNdebele.

Terminology plays a crucial role in lexicography as it shapes the precision and clarity with which dictionaries and other lexical resources convey meaning.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1926 (Article)

Due to the scarcity of language and lexicographic resources, general dictionaries could, according to Gouws and Prinsloo (2005: 61), include a broader selection of terms from clearly defined specialised fields. The treatment of such terms should be tailored to the layperson encountering them in everyday communication, rather than to experts in a specific field. In addition, the evolution of terminology reflects changes in language and society, influencing how new words are incorporated and defined. Thus, careful consideration of terminology is essential for maintaining the relevance and usefulness of lexicographical works.

Furthermore, the lack of standardised terminology has a detrimental effect on the growth of a language as this means that no new terms are created. Finlayson and Madiba (2002: 53) argue that for the terminology of a language to develop, intellectualisation must take place. They maintain that through this technique, underdeveloped African languages will develop more rapidly, and their terminology will carry the full weight of scientific rigour and clarity. In addition, they (ibid.) emphasise that intellectualisation ensures that language changes in a way that gives it the ability to carry and communicate all types of knowledge across all domains of life. The inclusion of terminology from well-defined specialised fields in general dictionaries ensures that the dictionaries reflect current and accurate language usage within a particular field. This approach is essential for keeping dictionaries up-to-date and relevant, particularly in rapidly evolving disciplines.

It is against this background that the current study examines the role of corpus-driven term extraction in filling the lexicographic gaps created by the lack of terms, with a particular focus on isiNdebele health terms. The research also establishes the success of the WordSmith Tools (WST) in identifying terms in a language with such a conjunctive orthography and the extent to which the WST reduces manual work.

In the next section an overview of similar research conducted in other African languages, will be reviewed. The focus will be on the types and sizes of corpora used, whether the corpora were written or spoken, and whether term extraction and analysis were done manually or with the assistance of software tools. This will be followed by a description of the resources that were used in this study. An exposition of the methodology used for term extraction of isiNdebele health terminology, a description of the results, and recommendations will be given. Finally, a conclusion and suggestions for future research will be presented.

2. Related work

The focus in this section is on semi-automatic term extraction conducted for other African languages. Of interest for this study are the types of tools used, the feasibility, practicality, and successes of various methodologies, emphasising the need for both computational and manual methods in terminological activities.

Taljard and De Schryver (2002) conducted a pioneering study on semi-automatic term extraction using basic corpus query software for African languages, particularly Sesotho sa Leboa. They employed three functions of the WordSmith Tools (WST): WordList, KeyWords, and Concord. Their findings revealed that the corpus query tool successfully identified 40% of multi-word linguistic terms that were overlooked manually. This led to the conclusion that semi-automatic term extraction significantly reduces human errors, proving to be both feasible and practical for African languages. Their findings also emphasised the fact that human beings will always remain the final judges in any terminological activity, whether that endeavour be manual or computational.

Building on the findings of Taljard and De Schryver, Prinsloo (2015) analysed corpora in Sesotho sa Leboa, English, and Afrikaans to assess the efficacy of restricted corpora for lexicographic endeavours. Using the Sketch Engine software, Prinsloo focused on frequencies and collocations, discovering that even lesser-resourced languages with limited, unbalanced corpora could yield results comparable to those of more resource-rich languages. This study challenged the notion of a "Big corpus", arguing that outcomes in languages with fewer resources such as the African languages, can be on par with those in languages with abundant resources.

Nkomo and Madiba (2011) employed semi-automatic term extraction to compile economics terminology for isiXhosa and Tshivenḓa. The study aimed to support concept literacy for students who are non-native English speakers. They used the WST and Multiconcord software, starting with the WordList function to generate a basic word list. After verification by economics lecturers, a final word list was created, and concordances were generated to identify term meanings in various contexts. This approach drew attention to the value of semi-automatic tools in educational settings.

Khumalo (2015) focused on the semi-automatic extraction of isiZulu linguistic terms with the goal of compiling dictionaries. Using both manual methods and the KeyWords function of the WST, Khumalo successfully extracted terms typical for the isiZulu linguistic domain. Terms identified by the tool were thereafter, manually verified. Khumalo (2015) outlined a term extraction process from raw corpora, without mentioning any efforts towards lemmatisation or morphological analysis to aid in term extraction. Notably, his article was published at a time when morphological analysers and lemmatisers for isiZulu were already accessible.

Ndhlovu (2014) investigated translation strategies for health terms from English to Zimbabwean isiNdebele using the parallel concordance tool, ParaConc. The English Ndebele Parallel Corpus (ENPC) was analysed to identify source terms and their equivalent translations in Zimbabwean Ndebele. ParaConc generated various data, including frequencies and potential translations, showing the tool's success. However, the researcher did not specify whether she manually verified the outcomes produced by the machine.

In a University of KwaZulu Natal (UKZN) project, a total of 1,863 terms was collected and subsequently deposited in the isiZulu term bank (Khumalo 2015:

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1926 (Article)

495-499). WST (version 6), primarily utilising the KeyWords function was applied for doing searches on full words. The isiZulu National Corpus grew significantly, and later Sketch Engine was used to create, manage and analyse the corpus (Khumalo 2018).

Mawonga et al. (2014: 66-68) report on the role of the South African-Norwegian Higher Education Development (SANTED) project in the development of African languages in various higher education institutions. This project focused on promoting multilingualism and developing indigenous South African languages and involved collaboration between institutions such as Rhodes University (RU), the University of KwaZulu-Natal (UKZN), and Durban University of Technology (DUT). UKZN and DUT developed an English–isiZulu term list and glossary for fields like education, nursing, and psychology. RU created multilingual resources across various disciplines. The project manually extracted and developed 1,400 terms, which were made available to nurses and midwives, enhancing language use in professional fields (Engelbrecht et al. 2010: 249-267).

The Special Language Corpora for African Languages (SPeLCAL) project aimed to develop linguistic resources for South Africa's nine official African languages, focusing on technical dictionaries, glossaries, and research in areas like terminology and translation. The project compiled written texts from fields such as politics, health, education, law, and technology. SPeLCAL also supported Second Language Teaching (SLT). The English–Venda Parallel Corpus pilot project used Multiconcord software for corpus extraction and analysis. Although it is not explicitly mentioned that terms identified by the software were manually verified, Madiba (2004) emphasises that in this English–Venda parallel corpus pilot project, it was discovered that small corpora allow early human intervention (Madiba 2004: 141, 146).

In summary, the above review of related work reveals varying approaches to term extraction, some relying solely on manual methods while others combine manual and tool-assisted approaches. None of the studies and projects mentioned the use of lemmatised or morphologically decomposed words, suggesting that only complete words were analysed. This review highlights the potential and challenges of semi-automatic term extraction in African languages, underscoring the importance of human judgment alongside computational tools.

3. Resources

In this section the resources employed in this study, namely written corpora and WST as an integrated suite of programs used for corpus analyses, are discussed.

3.1 Corpora

Two types of corpora were collected for this study namely, monolingual, written corpora for general purpose and monolingual, written corpora for language for specific purpose (health). Both corpora were collected from *Vuk'uzenzele*[2] newspapers and the South African Centre for Digital Language Resources (SADiLaR) repository. The *Vuk'uzenzele* newspapers can be accessed on the Government Communications Information Systems (GCIS) website. Published monthly, this paper is freely accessible and serves to keep South African citizens informed about government initiatives and services. It deals with a variety of topics with special focus on health, education, safety and security and rural developments. The topics are written in English and all other African languages including, isiNdebele. SADiLaR on the other hand, focuses on research and development in the domains of language-related studies and language technologies in the humanities and social sciences, for all of South Africa's official languages. Access to the repository is open to anyone interested in language technologies. Resources in the repository are categorised into two groups: downloadable and non-downloadable resources. Downloadable resources encompass various formats, such as sound recordings in MP3, portable document format (PDF) documents, plain text (.txt) files, and Microsoft (MS) Word documents. These formats are presented as they were received by SADiLaR. Individuals with research data in the fields of humanities, social sciences, and languages may submit their data to the SADiLaR repository. Consequently, medical corpora collected by Malele (2021), focusing on the use of corpora in compiling an English–isiNdebele glossary of medical terms, as well as the glossary created during this project, are now accessible on the SADiLaR repository[3].

The reference corpus (RC) used in this study is a language for general purpose corpus with 147 417 running words. The size of the RC ensures that a wide variety of subjects is covered, and that their content is diverse. The analysis corpus on the other hand is a domain-specific corpus with 99 052 running words. This corpus comprises of a variety of health topics, including TB, HIV, Primary Health Care, to mention but a few.

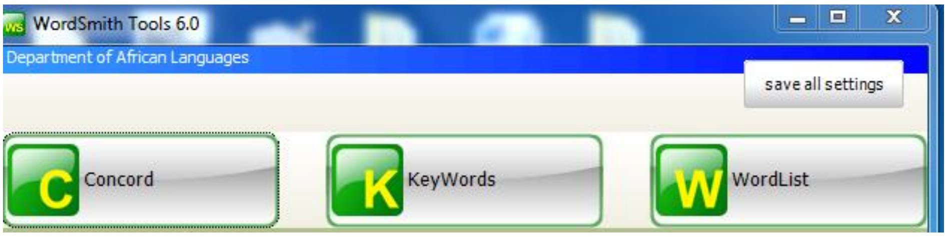


Figure 1: WordSmith Tools

3.2 WordSmith Tools

In this study, the WST tool (version 6.0) was employed. When reviewing literature where the same tool was used, the aim was to establish, among other aspects, the number of functions of the WST that various scholars had used. For

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1926 (Article)

instance, Nkomo and Madiba (2011) utilised only the WordList function and concordances of the WST, while Khumalo (2015) employed only the KeyWords function. Taljard and De Schryver (2002) made use of all the WST functions to extract and analyse Sesotho sa Leboa corpora. Similarly, the present study used all the functions of the WST to extract and analyse data from isiNdebele health corpora.

4. Methodology

The current research is rooted in the success of tried and tested methodologies employed in past projects and aims to refine and expand upon these proven approaches to further lexicographic research in African languages. Term extraction was conducted using a semi-automatic approach, combining both manual methods and the WST tool. The process involved utilising the tool to identify terms, which were then manually verified for accuracy. The following steps were followed:

Step 1: Collection of health and general texts: As mentioned earlier, both the isiNdebele health (monolingual, written and specialised) and general texts were collected from two sources namely, *Vuk'uzenzele* newspapers found on the website of the GCIS and also from SADiLaR's repository. All the texts (health and general) from *Vuk'uzenzele* newspapers were PDFs. From the SADiLaR platform, only health texts were collected. The health texts from SADiLaR were in two formats namely, PDF and plain text (.txt) format.

Step 2: Text conversion: All the texts were electronically collected and thereafter, text conversion for texts in PDF took place. The process of text conversion took place in two forms. Texts were firstly converted into MS Word. The purpose of this was to remove graphs, tables, and pictures before the process of loading data on WST could begin. After texts were converted into MS Word, the second step was to further convert the text into .txt format.

Step 3: Cleaning of the plain text format: The plain texts required manual cleaning, a task made easier by the .txt format. Cleaning is important as it effectively removes linguistic 'noise' which can arise from variations in grammar, structure, style, and clearly incorrect spelling in the use of basic language.

Step 4: Loading of corpus files on WST: After the converted texts were cleaned, the corpus files were then loaded on the WST for automatic term extraction. As illustrated in Figure 1, the WST has the following functions: Concord, KeyWords, and WordList all of which were used in identifying and analysing health terms from the given corpora.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1926 (Article)

5. Discussion

For this study, the available NCHLT Lemmatiser (2018) as well as the NCHLT Morphological Decomposer (2018) tools for isiNdebele as described by Eiselen and Puttkammer (2014) were applied to the relevant corpora. The aim was to execute term extraction based on lemmas and not merely on full words. However, it was found that both these NCHLT tools for isiNdebele are not suitable for this purpose due to the unreliable quality of the output, with the result that raw, unlemmatised corpora had to be used. Table 1 and Table 2 represent an excerpt of the experiment with the NCHLT tools.

Word	**Lemma**	**Comment**	**Expected Lemmatisation**
abodorhodere	abodorhodere	x	dorhodere
babodorhodere	babodorhodere	x	dorhodere
bodorhodere	bodorhodere	x	dorhodere
nabodorhodere	nabodorhodere	x	dorhodere
njengodorhodere	njengodorhodere	x	dorhodere
nodorhodere	nodorhodere	x	dorhodere
udorhodere	dorhodere	Lemma correctly identified	
yobudorhodere	yobudorhodere	x	dorhodere

Table 1: Results of lemmatisation with NCHLT Lemmatiser

An excerpt of the output of the NCHLT Lemmatiser (2018) in Table 2 indicates inconsistences in the sense that only the basic, singular word *udorhodere* 'doctor' is lemmatised correctly, whereas the plural form *abodorhodere* 'doctors' and other forms with possessive and adverbial prefixes are not lemmatised at all. Similarly, the accuracy of the NCHLT Morphological Decomposer (2018) output appears flawed, as demonstrated by the examples in Table 3. While the basic singular word *udorhodere* 'doctor' is correctly decomposed, the plural form *abodorhodere* 'doctors' is incorrectly decomposed, and all other forms with possessive and adverbial prefixes remain undecomposed.

Word	Decomposition	Comment	Expected Decomposition
abodorhodere	a-bodorhodere	x	abo-dorhodere
babodorhodere	babodorhodere	x	ba-bo-dorhodere
bodorhodere	bodorhodere	x	bo-dorhodere
nabodorhodere	nabodorhodere	x	na-bo-dorhodere
njengodorhodere	njengodorhodere	x	njenga-u-dorhodere
nodorhodere	nodorhodere	x	na-udorhodere
udorhodere	u-dorhodere	Morphological decomposition correct	
yobudorhodere	yobudorhodere	x	yo-bu-dorhodere

Table 2: Results of morphological decomposition with NCHLT Morphological Decomposer

Accurate lemmatisation or morphological decomposition for isiNdebele with its conjunctive orthography, could have impacted the frequency of word counts and reduced much of the manual work required for term identification. For example, the keyword *dorhodere* 'doctor' appears with a frequency of 75 in the Frequency Word List (see Table 3). If the lemma *dorhodere* had been correctly identified in the other words listed in Tables 1 and 2 using a lemmatiser or morphological decomposer, this lemma could have replaced the unlemmatised forms in the Frequency Word List, thereby significantly increasing the frequency of *dorhodere* to over 250.

a. WordList / Frequency Ranked List function

To begin, it was necessary to create word lists of health terms. The text/file icon was chosen, followed by the 'Make a word list now' icon. The output came in three different formats, namely the frequency-ranked word list, the statistical analysis and an alphabetically ordered word list.

The frequency WordList function was the first function to be used. It is worth noting that the tool was provided with the unlemmatised form of texts. All functions of the WST were utilised for this article, as they complement each other. Terms possibly overlooked by one function could be picked by another. The frequency WordList was used to extract synonyms and terms with variant forms. Here are the identified terms with variant forms:

*ingogwan**a*** vs *ingogwan**i*** 'virus'
*udorhoder**e*** vs *udorhoder**a*** 'doctor'
*umulwan**i*** vs *umulwan**a*** 'germ'

Table 3 represents an excerpt from the Frequency Word List containing unlemmatised health terms.

N	**Word**	**Frequency**	**%**	**Texts**	**%**
90	ubulwele	86,00	0,09	1,00	100,00
91	womnyaka	86,00	0,09	1,00	100,00
92	endaweni	85,00	0,09	1,00	100,00
93	imithetho	85,00	0,09	1,00	100,00
94	ihlelo	84,00	0,08	1,00	100,00
95	njengombana	84,00	0,08	1,00	100,00
96	sakho	84,00	0,08	1,00	100,00
97	ukusebenza	83,00	0,08	1,00	100,00
98	weengazi	83,00	0,08	1,00	100,00
99	izinga	82,00	0, 08	1,00	100,00
100	ukuze	82,00	0,08	1,00	100,00
101	zoke	82,00	0,08	1,00	100,00
102	angeze	80,00	0,08	1,00	100,00
103	nofana	80,00	0,08	1,00	100,00
104	of	80,00	0,08	1,00	100,00
105	lezamaphilo	79,00	0, 08	1,00	100,00
106	tb	79,00	0,08	1,00	100,00
107	inomboro	78,00	0,08	1,00	100,00
108	nezokuphepha	78,00	0,08	1,00	100,00
109	nje	78,00	0,08	1,00	100,00
110	ubujamo	78,00	0, 08	1,00	100,00
111	kanti	77,00	0,08	1,00	100,00

112	iindleko	76,00	0,08	1,00	100,00
113	dorhodere	75,00	0,08	1,00	100,00
114	esibhedlela	75,00	0, 08	1,00	100,00
115	njalo	75,00	0,08	1,00	100,00
116	abasebenzi	74,00	0,07	1,00	100,00

Table 3: Frequency WordList

Table 3 illustrates that health terms are few as compared to non-health terms, for instance ranked number 90 is the term *ubulwele* 'disease', with 86 occurrences. Ranked number 105 is the term *lezamaphilo* 'of health' with 79 occurrences. Ranked number 106 is the term *tb* 'tb' with 79 occurrences. Ranked number 113 is the term *dorhodere* 'doctor' with 75 occurrences, and ranked number 114 is the term *esibhedlela* 'at the hospital' also with 75 occurrences. The most frequent words in the analysis corpus are function or grammatical words such as *njengombana* 'as it is', *nofana* 'or', *kanti* 'whereas', *ukuze* 'so that', *nje* 'now' or 'this way', *njalo* 'always'. Function words in isiNdebele include word classes such as adverbs, conjunctions and pronouns which serve a grammatical purpose in a sentence, but typically carry little lexical meaning. Unlike function words, content words hold lexical significance and typically represent tangible or intangible entities, actions, attributes, or concepts. In isiNdebele, content words include nouns, verbs and adjectives. Table 3 comprises mainly nouns (such as *imithetho* 'rules', *ihlelo* 'plan', *abasebenzi* 'workers', *iindleko* 'costs', etc.), but also (auxiliary) verbs (e.g., *angeze* 'he can/may/might not come'). This serves as evidence that function words or grammatical words dominate frequency lists. To preserve content words, function words can be excluded from the word list.

b. KeyWords function

To make the key word list, both the reference (general corpus) and analysis (health) corpus files were uploaded on the KeyWords function of the WordSmith Tools. The non-language specific WST was therefore relied on with the KeyWords function being chosen followed by 'Make a keyword list now'. A list of key words was then produced. The purpose of using the KeyWords function was to extract all term candidates from the health corpus. It was used to calculate words which are key in a text, that is, words used much more frequently or much less frequently in each corpus. Through this function, terms used in the analysis (health) corpus were identified. Here the frequency of each word in the word list of health corpus was compared with the frequency of the same word in the reference word list. The output was a list of key words, or words whose frequencies are higher in the analysis corpus than in the RC. Any word which is found to be most outstanding in its frequency in the text is then considered to be 'key'. Key words are presented in their order of the most outstanding word. The KeyWords function provided the term candidate list as illustrated in Table 4.

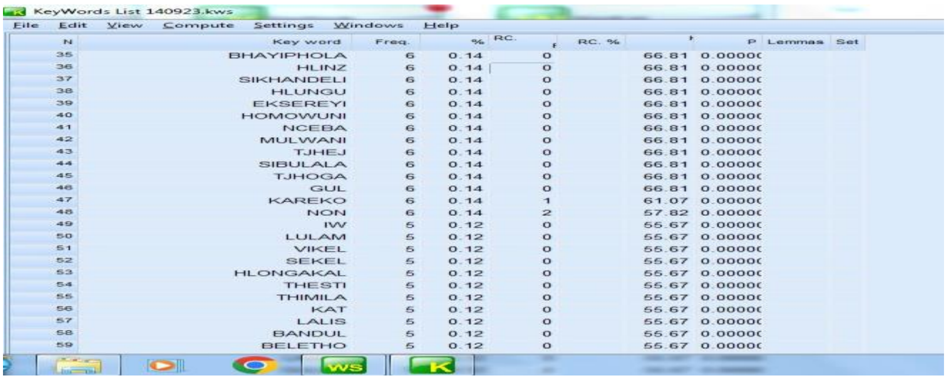
N	Key word	Freq.	%	RC.Freq.	RC. %	Keyness
1	mrhatjhi[4]	281,00	0,28	0,00		512,79
2	dorh	192,00	0,19	0,00		350,27
3	gems	188,00	0,83	0,00		342,97
4	begodu	822,00	0,45	498,00	0,34	262,45
5	khulu	448,00	0,16	187,00	0,13	239,87
6	HIV	154,00	0,19	13,00		202,92
7	ngamunye	184,00		28,00	0,02	198,92
8	amatshwayo	131,00	0,13	6,00		195,83
9	ingabe	169,00	0,17	26,00	0,02	181,84
10	umndeni	131,00	0,13	10,00		177,01
11	kobana	426,00	0,43	220,00	0,15	174,60
12	angabe	252,00	0,25	81,00	0,05	173,45
13	sista	93,00	0,09	0,00		169,61
14	beemali	122	0,12	12,00		154,04
15	umzimba	96,00	0,10	3,00		151,28
16	tb	79,00	0,08	0,00		144,07
17	womnyaka	86,00	0,09	2,00		139,80
18	dorhodere	75,00	0,08	0,00		136,77
19	lezamaphilo	79,00	0,08	1,00		134,34

20	ukudla	165,00	0,17	43,00	0,03	134,34
21	nezokuphepha	78,00	0,08	1,00		132,55
22	ngomkhawulo	70,00	0,07	0,00		127,65
23	esibhedlela	75,00	0,08	1,00		127,15
24	ubulwele	86,00	0,09	5,00		123,24
25	weengazi	83,00	0,08	4,00		123,02
26	udorhodere	72,00	0,07	1,00		121,76
27	iingazi	65,00	0,07	0,00		118,53
28	isana	69,00	0,07	1,00		116,37
29	emzimbeni	67,00	0,07	1,00		112,78
30	wesibhedlela	61,00	0,06	0,00		111,24

Table 4: First 30 words of the resultant key word list

The focus is on the health terms that appear in this first 30 list. Ranked number 2 is the word *dorh* 'doc' which is an abbreviation for *udorhodere* 'doctor'. Its frequency is 192 in the health corpus, and it is, zero (0.00) in the RC. Words which are key are 350.27. Ranked number 6 is the word HIV with the frequency of 154 in the health corpus, and the frequency of 13.00 in the RC. Its keyness is 202.92. Ranked number 19 is the word *lezamaphilo* 'of health' with the frequency of 79 in the health corpus, and the frequency of 1 in the RC. Its keyness is 134.34. Ranked number 24 is the word *ubulwelwe* 'disease' with the frequency of 86 in the health corpus and the frequency of 5 in the RC. It is 123.24 in keyness. Ranked number 25 is the word *weengazi* 'of the blood', with 83 frequencies in the health corpus and the frequency of 4 in the RC. It is 123.02 in keyness.

When considering the key words in Table 4, it is evident that most health terms, from this study's corpus have zero frequencies in the RC. This is evident in the case *of dorh* 'dr', *sista* 'sister', *TB* 'TB', *dorhodere* 'doctor', *iingazi* 'blood', *wesibhedlela* 'of the hospital'. It is important to note that the KeyWords function only managed to identify 138 health terms. The main challenge here is multiwords, that is strings of words (two or more words) that are considered to be one lexical unit. The KeyWords function neither identifies nor extracts multiwords. A multi-word such as *ikankere yomlomo wesibeletho* 'cervical cancer' appears as three separate words, with different ranks and frequencies. Therefore, only the word *ikankere* derived from the Afrikaans 'kanker' will appear as the health term. The word *umlomo* will just be translated as 'mouth' and *wesibeletho* will be translated as 'of the womb'. This means that the two words (*umlomo* and *wesibeletho*) will be rendered as non-health terms. Following the KeyWords function's identification of only 138 terms, we manually reviewed the word list of health terms and collected 582 health terms in total. This indicates that the KeyWords function overlooked many terms. Figure 2 clearly demonstrates the challenges associated with identifying multi-word terms.



N	Key word	Freq.	%	RC.	RC. %		P	Lemmas	Set
35	BHAYIPHOLA	6	0.14	0		66.81	0.0000		
36	HLINZ	6	0.14	0		66.81	0.0000		
37	SIKHANDELI	6	0.14	0		66.81	0.0000		
38	HLUNGU	6	0.14	0		66.81	0.0000		
39	EKSEREYI	6	0.14	0		66.81	0.0000		
40	HOMOWUNI	6	0.14	0		66.81	0.0000		
41	NCEBA	6	0.14	0		66.81	0.0000		
42	MULWANI	6	0.14	0		66.81	0.0000		
43	TJHEJ	6	0.14	0		66.81	0.0000		
44	SIBULALA	6	0.14	0		66.81	0.0000		
45	TJHOGA	6	0.14	0		66.81	0.0000		
46	GUL	6	0.14	0		66.81	0.0000		
47	KAREKO	6	0.14	1		61.07	0.0000		
48	NON	6	0.14	2		57.82	0.0000		
49	IW	5	0.12	0		55.67	0.0000		
50	LULAM	5	0.12	0		55.67	0.0000		
51	VIKEL	5	0.12	0		55.67	0.0000		
52	SEKEL	5	0.12	0		55.67	0.0000		
53	HLONGAKAL	5	0.12	0		55.67	0.0000		
54	THESTI	5	0.12	0		55.67	0.0000		
55	THIMILA	5	0.12	0		55.67	0.0000		
56	KAT	5	0.12	0		55.67	0.0000		
57	LALIS	5	0.12	0		55.67	0.0000		
58	BANDUL	5	0.12	0		55.67	0.0000		
59	BELETHO	5	0.12	0		55.67	0.0000		

Figure 2: Resulting KeyWords function on multi-words

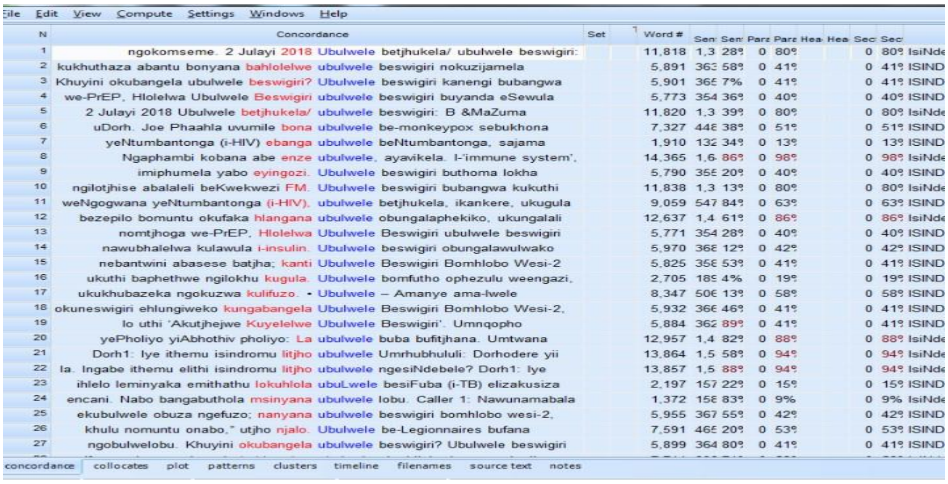
Figure 2 reflects the first version of the semi-automatically found term candidates. According to this figure, multi-words are not identified and extracted by the KeyWords function. Owing to the fact that the KeyWords function cannot sort the variants of the term candidate list, neither can it identify the multi-word term candidates, terms that resulted from the KeyWords function had to be manually validated. The researchers therefore had to manually sort certain terms and also manually identify multi-words. For instance, term number 37, *sikhandeli* loosely translates as 'preventer'. The full or complete word could be either *sikhandeli-magciwana* which loosely translates as 'preventer of germs', with 'antibiotic' as the correct health term, or *isikhandeli kuvuvuka* loosely translated as 'preventer of swollenness', with 'anti-inflammatory' being the correct health term. Term number 44 is also a multi-word of which the complete multi-word is actually *sibulala magciwana* loosely translated as 'the killer of germs'. The correct health term is 'antibiotic'. Figure 2 confirms that the KeyWords function indeed struggles to identify multi-word terms. Following this, we discuss the Concord function.

c. Concord function

To make a concordance, we started by choosing a text file. After a text file was

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1926 (Article)

chosen, search words were entered. Figure 3 shows the concordance for the search word *ubulwele* 'disease'.



N	Concordance	Set	Word #	Sen	Sen	Para	Para	Hea	Hea	Sec	Sec	
1	ngokomseme. 2 Julayi 2018 Ubulwele betjhukela/ ubulwele beswigiri:		11,818	1,3	28%	0	80%			0	80%	IsiNde
2	kukhuthaza abantu bonyana bahlolelwe ubulwele beswigiri nokuzijamela		5,891	363	58%	0	41%			0	41%	ISIND
3	Khuyini okubangela ubulwele beswigiri? Ubulwele beswigiri kanengi bubangwa		5,901	365	7%	0	41%			0	41%	ISIND
4	we-PrEP. Hlolelwa Ubulwele Beswigiri ubulwele beswigiri buyanda eSewula		5,773	354	36%	0	40%			0	40%	ISIND
5	2 Julayi 2018 Ubulwele betjhukela/ ubulwele beswigiri: B &MaZuma		11,820	1,3	39%	0	80%			0	80%	IsiNde
6	uDorh. Joe Phaahla uvumile bona ubulwele be-monkeypox sebukhona		7,327	448	38%	0	51%			0	51%	ISIND
7	yeNtumbantonga (i-HIV) ebanga ubulwele beNtumbantonga, sajama		1,910	132	34%	0	13%			0	13%	ISIND
8	Ngaphambi kobana abe enze ubulwele, ayavikela. I-'immune system',		14,365	1,6	86%	0	98%			0	98%	IsiNde
9	imiphumela yabo eyingozi. Ubulwele beswigiri buthoma lokha		5,790	355	20%	0	40%			0	40%	ISIND
10	ngilotjhise abalaleli beKwekwezi FM. Ubulwele beswigiri bubangwa kukuthi		11,838	1,3	13%	0	80%			0	80%	IsiNde
11	weNgogwana yeNtumbantonga (i-HIV), ubulwele betjhukela, ikankere, ukugula		9,059	547	84%	0	63%			0	63%	ISIND
12	bezepilo bomuntu okufaka hlangana ubulwele obungalaphekiko, ukungalali		12,637	1,4	61%	0	86%			0	86%	IsiNde
13	nomtjhoga we-PrEP. Hlolelwa Ubulwele Beswigiri ubulwele beswigiri		5,771	354	28%	0	40%			0	40%	ISIND
14	nawubhalelwa kulawula i-insulin. Ubulwele beswigiri obungalawulwako		5,970	368	12%	0	42%			0	42%	ISIND
15	nebantwini abasese batjha; kanti Ubulwele Beswigiri Bomhlobo Wesi-2		5,825	358	53%	0	41%			0	41%	ISIND
16	ukuthi baphethwe ngilokhu kugula. Ubulwele bomfutho ophezulu weengazi,		2,705	189	4%	0	19%			0	19%	ISIND
17	ukukhubazeka ngokuzwa kulifuzo. • Ubulwele – Amanye ama-lwele		8,347	506	13%	0	58%			0	58%	ISIND
18	okuneswigiri ehlungiweko kungabangela Ubulwele Beswigiri Bomhlobo Wesi-2,		5,932	366	46%	0	41%			0	41%	ISIND
19	lo uthi 'Akutjhejwe Kuyelelwe Ubulwele Beswigiri'. Umnqopho		5,884	362	89%	0	41%			0	41%	ISIND
20	yePholiyo yiAbhothiv pholiyo: La ubulwele buba bufitjhana. Umtwana		12,957	1,4	82%	0	88%			0	88%	IsiNde
21	Dorh1: Iye ithemu isindromu litjho ubulwele Umrhubhululi: Dorhodere yii		13,864	1,5	58%	0	94%			0	94%	IsiNde
22	la. Ingabe ithemu elithi isindromu litjho ubulwele ngesiNdebele? Dorh1: Iye		13,857	1,5	88%	0	94%			0	94%	IsiNde
23	ihlelo leminyaka emithathu lokuhlola ubuLwele besiFuba (i-TB) elizakusiza		2,197	157	22%	0	15%			0	15%	ISIND
24	encani. Nabo bangabuthola msinyana ubulwele lobu. Caller 1: Nawunamabala		1,372	158	83%	0	9%			0	9%	IsiNde
25	ekubulwele obuza ngefuzo; nanyana ubulwele beswigiri bomhlobo wesi-2,		5,955	367	55%	0	42%			0	42%	ISIND
26	khulu nomuntu onabo," utjho njalo. Ubulwele be-Legionnaires bufana		7,591	465	20%	0	53%			0	53%	ISIND
27	ngobulwelobu. Khuyini okubangela ubulwele beswigiri? Ubulwele beswigiri		5,899	364	80%	0	41%			0	41%	ISIND

Figure 3: Concord on search word *ubulwele* "disease"

Figure 3 clearly illustrates the concord function, and the role it played in deriving more terms. See the following lines:

Line 1 *ubulwele betjhukela*	'disease of sugar; sugar diabetes'
Line 2 *bahlolelwe ubulwele*	'they were tested for a disease'
Line 3 *okubangela ubulwele*	'causes the disease'
Line 6 *ubulwele be monkey-pox*	'disease of monkey-pox; Monkey-pox'
Line 7 *ubulwele bentumbantonga*	'disease of AIDS; AIDS'
Line 12 *ubulwele obungalaphekiko*	'incurable disease'
Line 13 *ubulwele beswigiri*	'disease of sugar; sugar diabetes'
Line 16 *ubulwele bomfutho ophezulu*	'disease of high blood pressure; high blood pressure'

Through the Concord function, term identification is streamlined, leading to the automatic discovery of more terms. For example, searching for the word *ubulwele* 'disease' yielded various types of diseases. In line 1 *ubulwele betjhukela* 'sugar diabetes', line 6, *ubulwele be-monkey pox* 'monkey-pox', line 7 *ubulwele bentumbantonga* 'AIDS', line 16, *ubulwele bomfutho ophezulu* 'high blood pressure'. All these disease types were identified through the use of the Concord search.

Utilising the Concord function has contributed to a deeper understanding of the usage of terms in context. It aided researchers in examining words within their textual context, facilitating the identification of patterns of similarity or contrast in the words surrounding the search term. For instance, the search term *ubulwele* 'disease' is frequently followed by the type of disease, as observed. More-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1926 (Article)

over, there appears to be a consistent pattern of words preceding the search term. In this instance, *ubulwele* 'disease' is preceded by *bahlolelwe* 'they were tested for' in line 2 *ebanga* 'that causes' in line 7. This suggests an association between the term 'disease' and terms such as 'tests', 'causes', and so forth.

Based on the given examples, the search term and its co-text are arranged so that the textual environment can be assessed and patterns surrounding the search term can be identified visually. Moreover, exploring concordances enables users to observe corpus occurrences, understand how meaning is constructed in texts, observe word co-occurrences, and recognise how they form meaningful patterns, without imposing predetermined notions on these units. As depicted in Figure 3 above, concordance analysis leads to the discovery of additional terms, simplifying the manual identification process. It also helps with collocates that is, the company that the key word keeps. Bowker and Pearson (2002: 124) state that collocates are words which typically occur in the vicinity of your search pattern. Collocates play an important role in corpus linguistics; they make it easier for the researcher or learner to understand the usage of words. Collocates further assist with the understanding of how two words come together meaningfully. Some concordances offer an additional facility which frequently ranks the words that appear in the vicinity of the search pattern.

Another significant aspect of the Concord function is its capability to identify multi-word expressions, a feature lacking in the Frequency WordList and KeyWords functions. Therefore, the Concord function produced the most effective results in terms of identifying both single and multi-word expressions. This is evident in the following lines:

Line 1: *ubulwele betjhukela/ubulwele beswigiri* 'diabetes' (loosely translated as 'disease of sugar')
Line 6: *ubulwele be-monkey pox* 'monkey pox' (loosely translated as 'disease of monkey pox')
Line 23: *ubulwele besifuba* 'TB' (loosely translated as 'disease of the chest')

The frequency WordList and KeyWords function would only identify *ubulwele* 'disease', *monkey-pox* 'monkey pox', *betjhukela* 'of sugar' and *besifuba* 'of the chest'.

In this section, the identification, extraction, and analyses of terms are discussed using the WST alongside manual methods. All the functions of the WST are fully utilised with manual verification of outcomes. Among these functions, the Concord function stands out for extracting more terms, including multi-word expressions not captured by the WordList and KeyWords functions.

Conclusion

Semi-automated term extraction in African languages yields valuable outcomes, particularly in addressing the deficiency of specialised terminology in isiNdebele. Terms extracted from specialised corpora or language for specific purpose (LSP) corpora, serve as an essential foundation for the compilation of subject field dictionaries by providing accurate and contextually relevant terms. This study has demonstrated the usefulness of semi-automatic term extraction in contributing

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1926 (Article)

to this foundation. While using software like WST in the processing of raw corpora to extract and identify single-word and multi-word terms, the bulk of terms required manual identification. However, manual identification does not negate the corpus-driven nature of the approach. Corpus-driven term extraction proves pivotal in mitigating terminology shortages, even in resource-scarce languages like isiNdebele.

The utilisation of WST in extracting and analysing data from available corpora underscores its potential applicability across various fields of study such as law, economics, and religion. It is apparent from this study that employing all functions of WST is crucial as these functions complement each other. Terms overlooked by one function may be captured by another. The Concord function, particularly adept at identifying multi-word terms, proves most fruitful compared to Frequency WordList and KeyWords functions.

Health terms identified in this study are intended for communication with laypersons rather than between experts in the health field. Consequently, given the scarcity of resources in the field of isiNdebele dictionaries, especially the lack of dictionaries for specific purposes, this range of health terms represents typical candidates for incorporation in a general dictionary. Terminology is vital in lexicography, ensuring reliable interpretation across dictionary entries. Additionally, as language and society evolve, terminology reflects these changes, influencing potential incorporation of new words. Therefore, careful consideration of terminology is crucial for maintaining relevant and useful lexicographical works.

The findings of this research hold significance for isiNdebele translators, lexicography students, educators, and linguists, offering insights into the role of technology in terminology resource development. Moreover, it contributes to the standardisation of health terms inconsistently used for communication in healthcare institutions.

Presently, due to the scarcity of isiNdebele corpora containing health-related terms, a broader range of general health terms had to be compiled. Future research endeavours will need to focus on utilising larger corpora to gather terminology in additional specialised healthcare fields.

Acknowledgements

The insights shared by Gertrud Faass during the co-supervision of the original doctoral research, the assistance provided by Marissa Griesel with the NCHLT Lemmatiser and Morphological Decomposer tools, and the constructive comments from two anonymous reviewers are gratefully acknowledged.

Endnotes

1. Southern Ndebele should be differentiated from Northern Ndebele (ISO 639-3: nde), spoken in Zimbabwe (cf. https://iso639-3.sil.org/)

2. https://www.vukuzenzele.gov.za
3. https://repo.sadilar.org/handle/20.500.12185/272
4. Although the first word with the highest frequency in Table 4 is a non-health term, namely *umrhatjhi* 'a radio DJ', it will be ignored for the purposes of this study because the relevant corpora were sourced from literacised radio interviews that all frequently included this term. This word has a frequency of 281.00 in the health corpus, and no frequency in the RC.

References

Bowker, L. and J. Pearson. 2002. *Working with Specialized Language: A Practical Guide to Using Corpora.* London: Routledge.

Eiselen, R. and M. Puttkammer. 2014. Developing Text Resources for Ten South African Languages. Calzolari, N. et al. (Eds.). 2014. *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC '14), Reykjavik, Iceland, May 26–31, 2014:* 3698-3703. Reykjavik, Iceland: European Language Resources Association (ELRA).

Engelbrecht, C., N.C. Shangase, S.J. Majeke, S.Z. Mthembu and Z.M. Zondi. 2010. IsiZulu Terminology Development in Nursing and Midwifery. *Alternation* 17(1): 249-272.

Finlayson, R. and M. Madiba. 2002. The Intellectualisation of the Indigenous Languages of South Africa: Challenges and Prospects. *Current Issues in Language Planning* 3(1): 40-61.

Gouws, R.H. and D.J. Prinsloo. 2005. *Principles and Practice of South African Lexicography*. Stellenbosch: SUN PReSS.

Jiyane, D.M. 1994. *Aspects of isiNdebele Grammar*. Unpublished M.A. Dissertation. Pretoria: University of Pretoria.

Khumalo, L. 2015. Semi-automatic Term Extraction for an isiZulu Linguistic Terms Dictionary. *Lexikos* 25: 495-506.

Khumalo, L. 2018. Towards an isiZulu National Corpus. Du Plessis, A.H. and S.E. Bosch (Eds.). 2018. *African Association for Lexicography, 23rd International Conference, June 27–29, 2018, University of the Western Cape, Cape Town, South Africa: Abstracts and Programme:* 26-29. Cape Town: AFRILEX.

Madiba, M. 2004. Parallel Corpora as Tools for Developing the Indigenous Languages of South Africa with Special Reference to Venda. *Language Matters* 35(1): 133-147.

Malele, N.J. 2021. *The Use of Corpora in the Compilation of a Specialised English–isiNdebele Glossary of Medical Terms.* Unpublished D.Litt. et Phil. Thesis. Pretoria: University of South Africa.

Mawonga, S., P. Maseko and D. Nkomo. 2014. The Centrality of Translation in the Development of African Languages for Use in South African Higher Education Institutions: A Case Study of a Political Science English–isiXhosa Glossary in a South African University. *Alternation Special Edition* 13: 55-79.

NCHLT isiNdebele Lemmatiser. 2018.
https://repo.sadilar.org/handle/20.500.12185/303 [15 April 2024].

NCHLT isiNdebele Morphological Decomposer. 2018.
https://repo.sadilar.org/handle/20.500.12185/304 [15 April 2024].

Ndhlovu, K. 2014. Term-creation Strategies Used by Ndebele Translators in Zimbabwe in the Health Sector: A Corpus-based Approach. *Stellenbosch Papers in Linguistics Plus* 43: 327-344.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1926 (Article)

Nkomo, D. and M. Madiba. 2011. The Compilation of Multilingual Concept Literacy Glossaries at the University of Cape Town: A Lexicographical Function Theoretical Approach**.** *Lexikos* 21: 144-168.

Prinsloo, D.J. 2015. Corpus-based Lexicography for Lesser-resourced Languages — Maximizing the Limited Corpus. *Lexikos* 25: 285-300.

SADiLaR. 2024. *South African Centre for Digital Language Resources.* https://repo.sadilar.org/handle/20.500.12185/272/ [25 March 2024].

Scott, M. 2010. WordSmith Tools (Version 5.0). [Computer software]. Liverpool: Lexical Analysis Software.

Taljard, E. and S.E. Bosch. 2006. A Comparison of Approaches to Word Class Tagging: Disjunctively vs. Conjunctively Written Bantu Languages. *Nordic Journal of African Studies* 15(4): 428-442.

Taljard, E. and G.-M. de Schryver. 2002. Semi-automatic Term Extraction for the African Languages, with Special Reference to Northern Sotho. *Lexikos* 12: 44-74.

Vuk'uzenzele newspaper English–isiNdebele. Government Communication & Information System (GCIS). https://www.vukuzenzele.gov.za [25 March 2024].

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1936 (Article)

Dictionaries in Context, Context in Dictionaries: Legal Translation Tools

Sandro Nielsen, *Aarhus University, School of Communication and Culture, Department of English, Denmark*
(sn@cc.au.dk) (https://orcid.org/0000-0002-8269-6800)

Abstract: Translators work with context and legal translation dictionaries may be tools that provide such context. However, lexicographers distinguish between different types of contexts, so it is relevant to examine which types of contexts are needed to help legal translators, bearing in mind that legal translation is an interdisciplinary activity involving competences and skills relating to law, language, and translation. Furthermore, legal translation involves a decoding, a transfer, and an encoding phase, each requiring different types of contexts from legal translation dictionaries. An examination of context in legal translation dictionaries treating the languages Danish, English, French, German, and Norwegian reveals that it may be necessary to distinguish between the context of dictionaries as information tools (dictionaries in context) and the context relating to the data they contain (context in dictionaries). Placing dictionaries in context concerns their format, size, scope, content, use, and user groups, while placing context in dictionaries concerns pragmatic contexts, syntactic-semantic contexts, and context of use related to source-language as well as target-language items, including concepts, terms, collocations, phrases, translation equivalents, example sentences, dictionary-internal cross-references, and dictionary-external references.

Keywords: LAW, TERMINOLOGY, PHRASEOLOGY, BILINGUAL DICTIONARIES, TRANSLATION DICTIONARIES, CONTEXTUAL DATA, CONTEXTUALIZATION, LEGAL LEXICOGRAPHY

Opsomming: Woordeboeke in konteks, konteks in woordeboeke: Hulpmiddels vir regsvertaling. Vertalers werk met konteks en regsvertalingswoordeboeke kan hulpmiddels wees wat sodanige konteks verskaf. Aangesien leksikograwe egter tussen verskillende tipes kontekste onderskei, is dit relevant om te ondersoek watter kontekste benodig word om regsvertalers van hulp te wees. Daar moet ook in ag geneem word dat regsvertaling 'n interdissiplinêre aktiwiteit is wat bevoegdhede en vaardighede rakende die regte, taal en vertaling behels. Voorts behels regsvertaling ook 'n dekoderings-, oordrag- en enkoderingsfase, wat elkeen verskillende tipes kontekste van regsvertalingswoordeboeke vereis. 'n Ondersoek na konteks in regsvertalingswoordeboeke waarin die tale Deens, Engels, Frans, Duits, en Noorweegs, hanteer word, bring aan die lig dat dit nodig mag wees om tussen die konteks van woordeboeke as inligtingshulpmiddels (woordeboeke in konteks) en die konteks wat verband hou met die data wat hulle bevat (konteks in woordeboeke) te onderskei. Om woordeboeke binne konteks te plaas, behels hul formaat, grootte, bestek, inhoud, gebruik en gebruikersgroepe, terwyl die plasing van konteks in woordeboeke die volgende behels: pragmatiese kontekste, sintakties-semantiese kontekste, en gebruikskonteks wat

Lexikos 34 (AFRILEX-reeks/series 34: 2024): 288-308

verband hou met die brontaal- sowel as doeltaalitems, insluitende konsepte, terme, kollokasies, frases, vertalingsekwivalente, voorbeeldsinne, woordeboek-interne kruisverwysings, en woordeboek-eksterne kruisverwysings.

Sleutelwoorde: REGTE, TERMINOLOGIE, FRASEOLOGIE, TWEETALIGE WOORDEBOEKE, VERTAALWOORDEBOEKE, KONTEKSTUELE DATA, KONTEKSTUALISERING, REGSLEKSIKOGRAFIE

1. Introduction

Translation depends on context as no translation occurs in a vacuum. In July 2022, the Council of the European Union issued a Decision whereby the EU acceded to the Convention on the Recognition and Enforcement of Foreign Judgments in Civil or Commercial Matters. This Decision and the Convention are available in all 24 official EU languages. The Danish translation of Article 2, point (g) of the Convention is interesting from both a translation and a contextual point of view. The original English text reads: "transboundary marine pollution, marine pollution in areas beyond national jurisdiction, ship-source marine pollution, limitation of liability for maritime claims, and general average" (Council of the European Union 2022). This point deals with maritime matters, the context of which is apparent from its wording, and the translation of the term *general average* into Danish is relevant as it has been rendered *generelt gennemsnit*. The maritime term *general average* refers to loss shared by a ship owner and a cargo owner caused by an act to save the ship or cargo (*The Oxford Companion to Ships and the Sea*), so this term falls within the maritime context. In contrast, the Danish translation corresponds to *arithmetic mean*, that is the result of adding two or more figures and dividing the result by the number of figures, and thus does not render the meaning of the English original. It would appear that the translators and post-editors of the Danish text lost sight of the relevant context. Reference works may be expected to help translators in types of translation situations like the one above, so it is relevant to examine whether, and if so to what extent and how, reference works can include contexts.

One of the challenges of including context in dictionaries is that the lexicographic community does not work with context as a single concept. The existing literature reveals that lexicographers mean and refer to different things when writing about context. An examination of the contributions in *The Routledge Handbook of Lexicography* (Fuertes-Olivera 2018) shows that lexicographers refer to many different types of context including, but not limited to, discursive context, factual context, sociolinguistic context, usage context, semantic context, collocational context, textual context, terminology context, communicative context, historic context, social context, extra-lexicographical context, linguistic context, intercultural context, syntactical context, interprofessional context, associative context, knowledge-rich context, and non-linguistic context. So, what do lexicographers mean when they refer to context in connection with practical and theoretical lexicography?

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1936 (Article)

One way of providing an answer is to study context in relation to dictionaries intended for translation of legal texts. When discussing context in legal dictionaries, lexicographers should treat law as a jurisdiction-dependent domain, as De Groot (1990: 122) explains: "Die Fachsprache der Juristen ist extrem systemgebunden. Rechtssysteme sind von Staat zu Staat unterschiedlich. Jeder Staat hat seine eigene selbständige juristische Terminologie" [The jargon used by lawyers is extremely system-bound. Legal systems vary from state to state. Each state has its own independent legal terminology]. The challenge of legal translation is further accentuated by Biel (2022: 379): "Because it requires cross-linguistic mediation in the field of law, legal translation is an interdiscipline at the intersection of translation studies, legal studies, linguistics and terminology." With direct reference to legal translation dictionaries, Prieto Ramos (2021: 178) sums up the needs of legal translators: "Bilingual legal dictionaries, in particular, have been traditionally regarded in the field as resources of limited reliability due to insufficient contextualization and comparative legal information for translation decision-making." In other words, each jurisdiction structures its legal system in a way that suits only that jurisdiction and the legal language used reflects the structure of that system, thereby establishing its own context.

This article will discuss context in legal translation dictionaries and how contextual data can help legal translators in the translation process. This will involve two Scandinavian languages and English, French and German. The three world languages are often used in international, legal communication, while Danish and Norwegian are the small, Scandinavian languages chosen, because legal terms in the two languages are often very similar, in contrast to Swedish and Finnish legal terms. This article also discusses the context in which dictionaries are embedded and its implications (Section 2) and then examines contexts related to the data contained in legal translation dictionaries (Section 3). Finally, the relation between dictionary-internal contexts and the legal translation process is examined (Section 4).

2. Placing dictionaries in context

Not only do dictionaries treat words, terms, language, and facts that are rooted in different contexts, but dictionaries are themselves rooted in contexts. There is no official definition of *dictionary*, but the one given in *Dictionary of Lexicography* seems to be a good place to start the search for those contexts. Hartmann and James (2001: 41) provides the following definition of *dictionary*: "A type of REFERENCE WORK which presents the vocabulary of a language in alphabetic order, usually with explanations of meanings" (emphasis in original). This definition represents what may be called entities, i.e. individual, real objects within the domain of lexicography, and they are instances of a class or type. According to Bergman (2018: 132), classes and types are not individual, real objects but constructs of thought. Following this line of research, a further examination of Hartmann and James (2001) shows that dictionaries come in many

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1936 (Article)

varieties, such as abridged dictionaries, defining dictionaries, electronic dictionaries, frequency dictionaries, general dictionaries, historical dictionaries, learner's dictionaries, paper dictionaries, technical dictionaries, and translation dictionaries. This short list contains varieties of dictionaries and shows how different they are from each other; they vary in format, size, scope, content, use, and user group. The label *dictionary* is not a very informative example of the concept of lexicographic products, but the referential information contained in the adjectives and nouns in the above list helps placing dictionary variants in their proper contexts.

Legal translation dictionaries can be described as needs-adapted information tools intended to provide help to translators of legal texts. Lexicographers may therefore find it challenging to identify and describe the contexts in which their dictionaries are embedded as well as the needs of dictionary users. The term *context* is found in many different domains and a search for the term in the dictionaries included in *Oxford Reference Online* (2024) reveals that context is defined in dictionaries of disciplines such as archaeology, biochemistry, communication science, epidemiology, geography, linguistics, media science, occupational science, and philosophy, and every time with domain-specific meanings so that the concept meets their own domain's explanatory needs. There does not appear to be a specific meaning of *context* within the field of law, though contexts are regularly referred to in legal texts such as contracts, books, statutes, and judicial rulings involving legal interpretation (Bajčić 2017: 74-78). Translators distinguish between two broad types of contexts, namely text-internal contexts, which comprise various context types at different textual levels, and text-external contexts, which may refer to any relevant concepts outside the texts to be translated, a distinction that corresponds to what Sandrini (2018: 558-559) calls the textual universe and the legal universe. Such texts, including legal texts, often refer to tangible and intangible objects or concepts that belong to one of more domains. In contrast, Zgusta (1971: 227) specifies lexicographic context as restricted to lexical dimensions, i.e. verbal context, which seems to cover only some of the types of contexts relevant for lexicography identified above, thereby confirming the elusive nature of the term *context* in lexicography.

Dictionaries may be regarded as vehicles of knowledge representation in that they contain data that users can process to gain information and knowledge about parts of the world. In this light, knowledge representation studies may provide a definition that is suitable for establishing a context for dictionaries and one such definition is suggested by Baclawski et al. (2018: 182): "In general, a context is commonly understood to be the circumstances that form the setting for an event, statement, process, or idea, and in terms of which the event, statement, process, or idea can be understood and assessed." One interpretation is that this definition refers to context as something external to a given point of departure, e.g., external to dictionaries. Since dictionaries are classified and function as information tools, it may be appropriate to distinguish between the context of dictionaries as information tools (dictionaries in context) and the context relating to the data they contain (context in dictionaries).

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1936 (Article)

At an overall level, the context can be regarded as the circumstances surrounding the dictionary from the planning stage through to the use stage. This would imply that context is introduced almost as a sort of "super context" that relates to the dictionary as a fundamental basis for the decisions that have to be taken during the planning, designing, completion, and use stages. In an attempt to identify the elements of such an overall context of legal translation dictionaries, the point of departure will be the variance of dictionaries mentioned above so that the context of dictionaries will include at least the following elements: format, size, scope, content, use, and user group. It is beyond the scope of this article to analyse all these components in detail, and the following discussion focusses on some of the issues directly relevant for discussing contexts surrounding dictionaries as information tools.

The first element concerns the *format* in which dictionaries are designed and published. Formats may include paper dictionaries, digital dictionaries, offline dictionaries, and online dictionaries, and the type of format affects the editorial work as well as the presentation of and access to data. Today most dictionary projects use digital means for data extraction and use databases for general editorial work whether the dictionaries are published in print or online. One area in which the difference between printed and digital dictionaries is most pronounced is the presentation of data in that many presentational options are available online, e.g. use of links, colours, pop-up text boxes, video footage, oral presentation of articles, and other audio-visual means. In addition, internet browsers such as Google Chrome and Microsoft Edge include functionalities that allow users to have dictionary articles read aloud as well as translated into other languages.

Size refers to the physical and digital dimensions of dictionaries and usually includes such issues as the number of words and terms in information tools, and the number of print pages and webpages. It is still reasonable to distinguish between small and large printed dictionaries, while online dictionaries may be characterised as either data-rich or data-poor when it comes to comparing dictionaries. Even though databases allow lexicographers to have large volumes of data for lexicographic projects, the way in which such data are presented to users may be limited in order to avoid information overload and to accommodate users with small-screen devices (see e.g. Lew 2011).

The third element of overall context is *scope*, which generally refers to the relative specialisation of dictionaries, such as "the range of topics treated, the technicality of details included and the timescale covered" (Hartmann and James 2001: 122). The scope of legal translation dictionaries covers law, legal language and translation. As suggested by Nielsen (1990: 132-135), it is possible to distinguish between general-field dictionaries, which are intended to cover the entire field of law, and sub-field dictionaries, which are intended to cover a specific sub-field of law, e.g. company law and land law. This distinction concerns both lexicographers and dictionary users because it affects the data to be selected and the help dictionaries can provide. General-field dictionaries can potentially contain more words and terms than sub-field dictionaries and their possibility

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1936 (Article)

of use is greater than that of sub-field dictionaries as they cover the entire subject-field of law. On the other hand, sub-field dictionaries can potentially treat the sub-fields concerned more extensively and detailed than general field dictionaries, especially printed ones. This means that sub-field dictionaries can contain data that provide help in respect of legal facts/rules, legal vocabulary, derivation, grammatical irregularities, definitions, translation equivalents, degrees of equivalence, translation of collocations, translation of phrases, textual conventions in legal genres and sub-genres, and translation strategies.

For the purposes of this article, *content* is the data contained in a dictionary and lexicographers often distinguish between various types of content. Firstly, function-related data are the data that support the functions for which dictionaries are intended, e.g. providing help to translate legal terms, and this type of data is mostly found in dictionary articles but may also be found in various appendices. Secondly, lexicographers distinguish between textual data, visual data, and audio data, a distinction that concerns the way in which editors collect, store, and maintain data, the way in which data are presented to users, as well as the way in which users can access the data. Thirdly, data content is directly related to the subject field(s) treated by dictionaries, e.g. the legal domain in single-field dictionaries and sub-field dictionaries. Finally, content is related to the size of dictionaries, and it is possible to distinguish between data-richness and data-poverty, e.g., when reviewing dictionaries with focus on the presentation and definition of words and terms, the number of collocations, phrases, and example sentences, as well as the help provided to translate legal texts.

In the fifth place, *use* refers to the types of situations in which people consult dictionaries, which are related to the types of help (i.e., functions) dictionaries are intended to provide. As indicated above, dictionaries are tools of information, and the meaning of this is explained in one of the contributions to *The Routledge Handbook of Lexicography*, which states that dictionaries are utility tools that are "conceived for consultation with the genuine purpose of meeting punctual information needs experienced by specific types of potential users in specific types of extra-lexicographical context" (Tarp 2018: 246). Extra-lexicographic contexts are examined and analysed by several researchers in the *Handbook* and include text reception, text production, translation, and LSP contexts. Lexicographers study the human activities their dictionaries are intended to address in order to identify general types of use situations. These activities may be divided into distinct types of actual situations in which users encounter problems that cause them to consult dictionaries in search for help. In addition, use situations are events that have nothing to do with lexicography and dictionaries, but which may lead to dictionary consultation: Actual use situations arise in extra-lexicographic environments and prior to actual dictionary consultation, e.g., when translating legal texts. This description of use supports the idea of working with extra-lexicographic contexts as well as dictionary-data contexts.

The last element of the overall context is the *user group*, which refers to the intended group of people for whom lexicographers have decided to provide help. To accommodate user needs lexicographers examine which competences

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1936 (Article)

particular types of users are likely to have, and depending on the help dictionaries are intended to give, the following user competences may be considered: native general-language competence; foreign general-language competence; cultural competence; factual competence; native special-language competence; foreign special-language competence; native-language text production competence; foreign-language text production competence; translation competence (see e.g. Nielsen 2023: 435-438). When these competences have been identified, lexicographers will have an indication of which types of data users need in specific types of use situations, because users need dictionaries to help them where their own competences are insufficient. As a result, lexicographers attempt to select data that can fill the competence lacunae and thereby provide the help that users require.

Users of legal translation dictionaries can be described in terms of factual competence, and Bergenholtz and Kaufmann (1997: 98-99) distinguish between laypersons, semi-experts, and experts in relation to the subject field in question. However, law is a culture-bound subject field with culture-related concepts and systematic structures. Consequently, experts may be specialists in their own culture but not in the corresponding field of law within another culture, which affects dictionaries treating two or more cultures and hence languages because users have different levels of subject-specific knowledge in the relevant cultures.

Secondly, users have different levels of language competence, and this distinction implies that users have from little to considerable linguistic competence in their native language, and perhaps in a foreign language. However, Fluck (1985: 12) argues that the language of a subject field (LSP) is characterised by using linguistic structures and options that are either more or less frequent than in the corresponding general language. This means that users of legal translation dictionaries may be expected to have limited knowledge of the relevant degrees to which certain linguistic structures and other options are used in legal texts in their native language as well as a foreign language. There may thus be a significant asymmetry of factual and language competences among users of legal translation dictionaries.

One aspect of context that the above list does not include is the dictionary project staff and their competences and skills. Nielsen (2018) examines the interdisciplinarity of lexicography and argues that lexicography involves cooperation between several disciplines. Firstly, IT specialists contribute to the development of databases and user interfaces together with lexicographers. Secondly, many dictionary projects are based on electronic corpora requiring knowledge about corpus building and analysis. Thirdly, general and specialised dictionaries tend to contain words and terms from several subject fields, so lexicographers need domain-specific knowledge and may have to work with domain experts to establish the internal structure of subject fields as well as the language used. Finally, dictionaries attempt to fulfil the information needs of users in actual situations unrelated to lexicography such as text production, text reception, translation, and copy-editing. Therefore, lexicographers may decide

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1936 (Article)

to cooperate with specialists from and to acquire knowledge about various disciplines so that, for example, linguists assist with linguistic data, domain specialists assist with factual data, and translation specialists provide data relevant for translation. The competences and skills of dictionary project staff are thus related to the other contextual elements: format, scope, content, use, and user group.

Against this background, the overall context of legal translation dictionaries may generally be described as the circumstances that form the setting for a dictionary or dictionary project in terms of the format, size, scope, content, use, user group, and dictionary project staff, and in relation to which that dictionary or project can be understood and assessed. Moreover, the discussion above indicates that several of these contextual elements may affect the contextual relationships concerning the data in dictionaries. The following section examines some of the types of data-related contexts with examples from legal translation dictionaries.

3. Placing context in dictionaries

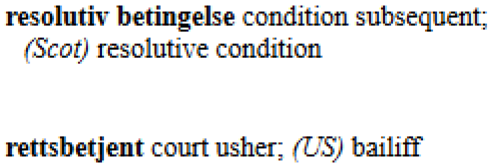
When discussing context in dictionaries, lexicographers face the same challenge of defining context as in the case of dictionaries in context. One reason for the many faces of context is the background of lexicographers, who have different special-domain backgrounds (based on different domain-specific theories and methodologies) linked to the interdisciplinarity of lexicography. This state of affairs is aptly described by Baclawski et al. (2018: 184): "It seems that, while context is important, all one can say in general about a context is that it is at a metalevel relative to the subject matter and that the context affects the interpretation of the subject matter." However, as indicated above, lexicographers work with many different types of contexts, which Kipfer (2022: 443) describes as follows:

> Context: a phrase, sentence, or paragraph surrounding a lexical unit that depicts its meaning or sense; also called lexicographic context, minimal context, situational context, context of use. Taken from either written or spoken sources, context shows the characteristic features of a lexical unit and the setting or circumstances with which a word or phrase is associated.

Applied to legal lexicography, this means that context in legal translation dictionaries would be phrases, sentences or paragraphs surrounding legal terms and concepts that depict their meaning or sense. Since Kipfer's definition is a general one, it may be appropriate to add that context in legal translation dictionaries may also have to include the extra-lexicographic translation process as well as the competences and skills of translators.

An examination of context in legal translation dictionaries would benefit from a more specific explanation of context in lexicography. Domínguez Vázquez and Gouws (2023: 236) suggest that context may be seen as a multidimensional concept covering "cotext, contextualisation, dictionary-internal context, dictionary-external context, and external dictionary context". This means that legal translation dictionaries should present legal terms and concepts with examples of their occurrence, include language use in the legal world and the world of legal translation, and refer or link to dictionary-external sources. Consequently, pragmatic and syntactic-semantic contexts as well as contexts of use are relevant and may relate to the source-language items as well as the target-language items, such as legal terms and concepts, collocations, phrases, and translation equivalents.

A traditional way of providing context in legal translation dictionaries, and in dictionaries generally, is to use labels. These are descriptors that indicate restricted use related to, e.g., domains or language usage and may be addressed to lemmas (i.e., entry words), definitions, equivalents, collocations, phrases, and example sentences. Diatopical labels, which indicate regional or dialectical restrictions, are often found in legal translation dictionaries, and are normally addressed to equivalents when intended to help translating into a foreign language, as illustrated in Figure 1.



resolutiv betingelse condition subsequent;
(Scot) resolutive condition

rettsbetjent court usher; *(US)* bailiff

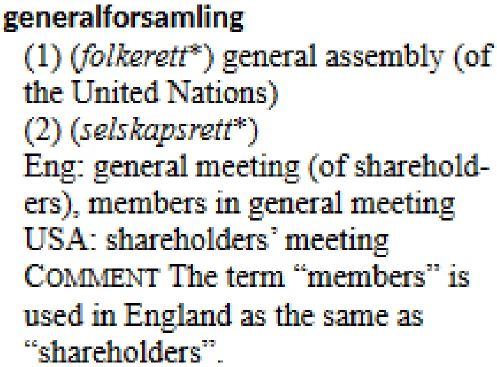
Figure 1: Diatopical labelling in articles from *Norsk–engelsk juridisk ordbok*

The articles in Figure 1 are from a Norwegian–English dictionary of law that has translators, interpreters, lawyers, and others who need to transpose Norwegian legal terminology to English as its intended user group. The labels, presented in abbreviated form (*Scot* for Scotland, and *US* for USA), help translators to select the correct equivalent in a given situation by indicating in which contexts they may be appropriate. The purpose of diatopical labels as context markers in legal translation dictionaries is twofold: They indicate which regional language variant equivalents belong to and, perhaps more importantly, which legal jurisdiction equivalents belong to. This duality is directly linked to the culture-bound nature of legal jurisdictions, concepts, and terms as well as their languages. Dictionaries translating into English present lexicographers with the challenge that English has many variants depending on the legal jurisdictions that form the context of equivalents, and *Norsk–engelsk juridisk ordbok* uses 8 diatopical labels relating to the following jurisdictions: America (USA), Australia, England and Wales, European Union, Ireland, Norway, Northern Ireland, and Scotland.

Diatechnical labels are related to the scope of dictionaries and indicate domain-specific contexts and their languages and may place lemmas and equiv-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1936 (Article)

alents in domains such as finance, medicine, and zoology. Since legal translation dictionaries are rooted in the domain of law, the data in such dictionaries are implicitly placed in a context of law. Though this is true, the above discussion indicated that the domain of law may be divided into sub-fields, which have to be explicitly presented in, e.g., general field dictionaries that treat the entire legal domain. The use of diatechnical labels in a general-field dictionary of law is illustrated in Figure 2.

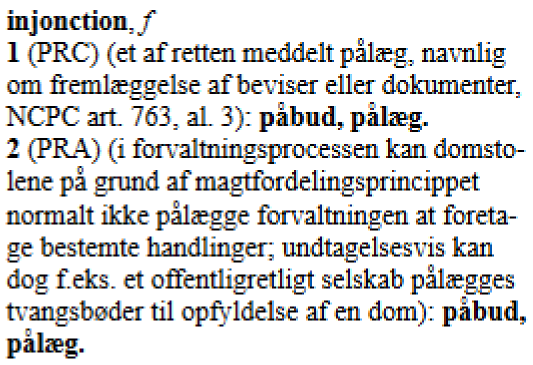


generalforsamling
(1) (*folkerett**) general assembly (of the United Nations)
(2) (*selskapsrett**)
Eng: general meeting (of shareholders), members in general meeting
USA: shareholders' meeting
COMMENT The term "members" is used in England as the same as "shareholders".

Figure 2: Diatechnical labels in *Stor norsk–engelsk juridisk ordbok*

The diatechnical labels found in Figure 2 show that legal dictionaries may have to present contexts that are specific for different sub-fields of law. The article is from a Norwegian–English law dictionary intended for Norwegian lawyers, translators, businesspeople, teachers, and public administrators who need to communicate about Norwegian legal matters in English. Figure 2 explains that the translation of the Norwegian term (or rather two concepts) *generalforsamling* should be translated according to the sub-field context in which it/they occur. The first context is indicated as *folkerett* (i.e. public international law) and the other as *selskapsrett* (i.e., company law) so that translators can select the appropriate English equivalent. At the same time, the article contains the diatopic labels *Eng* and *USA* to make translators explicitly aware of restrictions in the context of use. Finally, Figure 2 includes a comment informing users of two synonyms used in legal contexts, a comment that is addressed to the context labelled company law.

Sub-field dictionaries within the field of law treat a specific part of the legal domain, meaning that the functional data in the dictionaries are clearly rooted in an implied context. Nevertheless, it may be necessary to present context explicitly because sub-field dictionaries often provide an in-depth treatment of a sub-field and its terms, concepts, and associated language. Figure 3 illustrates how this may be done.

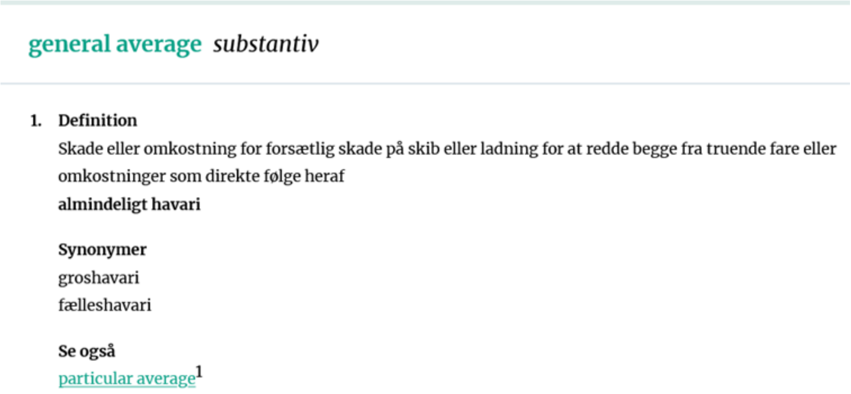


injonction, *f*
1 (PRC) (et af retten meddelt pålæg, navnlig om fremlæggelse af beviser eller dokumenter, NCPC art. 763, al. 3): **påbud, pålæg.**
2 (PRA) (i forvaltningsprocessen kan domstolene på grund af magtfordelingsprincippet normalt ikke pålægge forvaltningen at foretage bestemte handlinger; undtagelsesvis kan dog f.eks. et offentligretligt selskab pålægges tvangsbøder til opfyldelse af en dom): **påbud, pålæg.**

Figure 3: Diatechnical labels in a sub-field dictionary treating French and Danish: *Retsplejeordbog*

Figure 3 shows an article from the French–Danish section of a quadrilingual judicial dictionary that treats the sub-field law of procedure (i.e. the administration of justice), covering terms in English, French and German-speaking jurisdictions in the European Union and their Danish translations. Procedural law is often divided into two, or in the case of France three, broad sub-sub-fields, namely civil procedure, criminal procedure, and administrative procedure. The first context in which the French *injonction* occurs is civil procedure (indicated by PRC = *procédure civile*), whereas the second context is administrative procedure (indicated by PRA = *procédure administrative*). In this article, the context markers are addressed to the lemma and, in addition to the abbreviated labels, the contexts are provided by explicit definitions in Danish to cater for the intended Danish users. The definitions provide knowledge-related context and the boldface, Danish equivalents provide translation assistance. Finally, context is also provided by the reference to a source outside the dictionary (NCPC art. 763, al. 3), which is a reference to Article 763, point 3 of the *Nouveau Code de Procédure Civile*, where further details and context may be found. Another example of context in dictionaries appears in Figure 4.

The article treating the term *general average* is from an English–Danish law dictionary intended for Danish translators, lawyers and students who need to translate English legal texts into Danish (Figure 4). The Danish definition explains the meaning of the English term and places it in a maritime context, matching the above definition in *The Oxford Companion to Ships and the Sea*. Furthermore, the article presents a recommended Danish translation in boldface, *almindeligt havari*, as well as two synonyms (*groshavari*, *fælleshavari*) to the Danish equivalent. In legal translation dictionaries, it is important that equivalent-addressed synonyms are full synonyms (as is the case in Figure 4), i.e. have the same meaning in law and are not subject to, e.g., collocational restrictions. Finally, a cross-reference to another article in the dictionary (*particular average*) provides additional context, allowing users to compare and contrast the two types of average in maritime law.



general average *substantiv*

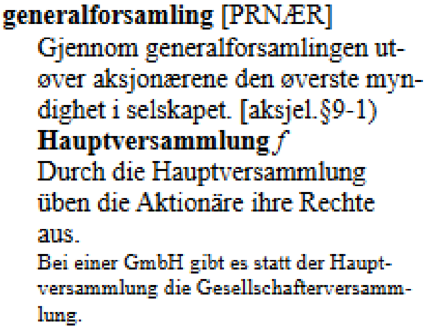
1. **Definition**
Skade eller omkostning for forsætlig skade på skib eller ladning for at redde begge fra truende fare eller omkostninger som direkte følge heraf
almindeligt havari

Synonymer
groshavari
fælleshavari

Se også
particular average[1]

Figure 4: Context presented as definition in *Juridisk ordbog engelsk–dansk/dansk–engelsk*

Lexicographers may provide extended context by presenting definitions in both the source and the target language. This type of contextual presentation may assist translators in two ways: the source-language definition helps users to establish whether they have found the correct dictionary article to check that the meaning of the lemma matches that in the source text; secondly, the target-language definition may help users with translation-relevant context in that it helps users translate or otherwise communicate about a source-language term in the target language. One example of this type of context is found in Figure 5, which contains an article from a Norwegian–German dictionary of law with professional translators, lawyers, and students as its intended user group.

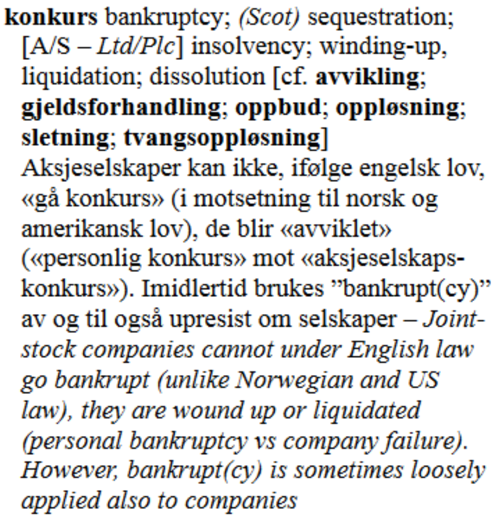


generalforsamling [PRNÆR]
Gjennom generalforsamlingen utøver aksjonærene den øverste myndighet i selskapet. [aksjel.§9-1)
Hauptversammlung *f*
Durch die Hauptversammlung üben die Aktionäre ihre Rechte aus.
Bei einer GmbH gibt es statt der Hauptversammlung die Gesellschafterversammlung.

Figure 5: Context presented in definition in two languages: *Norsk–Tysk Juridisk Ordbok*

The two definitions in Figure 5 are directly related to the overall contexts of scope, function, and use. The label PRNÆR indicates that the Norwegian term belongs to the sub-field of commercial law. The Norwegian definition closes with a reference to a dictionary-external source, namely section 9-1 of the *Norwegian Companies Act*, where users can find additional context (the Act has subsequently been repealed and replaced). Since the dictionary provides translating assistance, the German equivalent and the German definition represent data that relate to the dictionary-function context as well as the use context, while the note in small print at the foot of the article makes users aware of differences in legal systems: the Norwegian term corresponds to a different German term if it relates to one specific type of company in Germany, namely a *Gesellschaft mit beschränkter Haftung* (GmbH) —– which may be described as a small or medium-sized private limited-liability company.

Legal translation dictionaries may also contain bilingual presentations of contextual data relating to differences in the two legal systems involved. This type of contextual data may take various forms as illustrated in Figure 6.

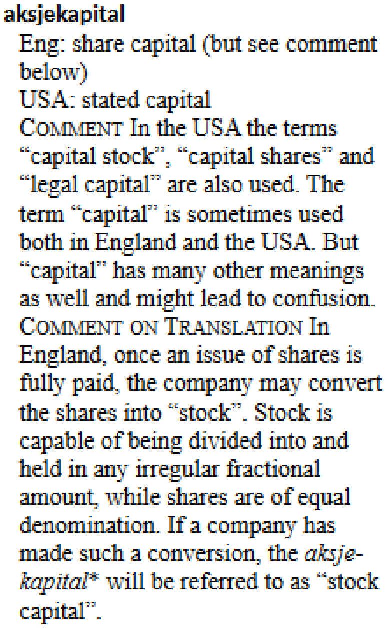


konkurs bankruptcy; *(Scot)* sequestration; [A/S – *Ltd/Plc*] insolvency; winding-up, liquidation; dissolution [cf. **avvikling**; **gjeldsforhandling**; **oppbud**; **oppløsning**; **sletning**; **tvangsoppløsning**]
Aksjeselskaper kan ikke, ifølge engelsk lov, «gå konkurs» (i motsetning til norsk og amerikansk lov), de blir «avviklet» («personlig konkurs» mot «aksjeselskaps-konkurs»). Imidlertid brukes "bankrupt(cy)" av og til også upresist om selskaper – *Joint-stock companies cannot under English law go bankrupt (unlike Norwegian and US law), they are wound up or liquidated (personal bankruptcy vs company failure). However, bankrupt(cy) is sometimes loosely applied also to companies*

Figure 6: Excerpt from *Norsk–engelsk juridisk ordbok*

The article in Figure 6 is intended to help translators, interpreters, lawyers, and others translate Norwegian legal texts into English and the lexicographer has opted for presenting context that matches the profiles of the intended users. Firstly, the article contains several English equivalents: an unmarked equivalent (*bankruptcy*); a Scottish equivalent (*sequestration*) marked by a diatopical label; and four equivalents that are labelled as belonging to companies (*A/S* and *Ltd/Plc*). The interpretation of the label *Ltd/Plc* requires some knowledge on the part users as it refers to two corporate entities in the UK jurisdiction. Secondly, the equivalents are followed by cross-references in square brackets to six Norwegian terms where users may find additional contextual and translation-relevant data. Thirdly, the article explains in Norwegian that the term *konkurs* applies to both personal insolvency and company insolvency in contrast to the UK legal system, where the rules and terminology relating to personal and company insolvency differ. The legal-systemic data also explains that the Norwegian concept of *konkurs* is more like the US legal system. In addition, the English term *bankruptcy* is often loosely used to refer to both personal and company insolvency. Finally, the contextual data on legal-systemic differences is presented in both Norwegian and English, which may help users understand the differences as well as help users communicate about these differences in English. The contextual data in the article are thus related to the overall contexts of scope, function, and use.

Some dictionaries contain very specific data intended to assist legal translators. Such data are directly related to the context of use and may be presented in various ways, for instance, depending on the perspective adopted by lexicographers or related to translation strategies. Figure 7 contains an example of a comment on translation.



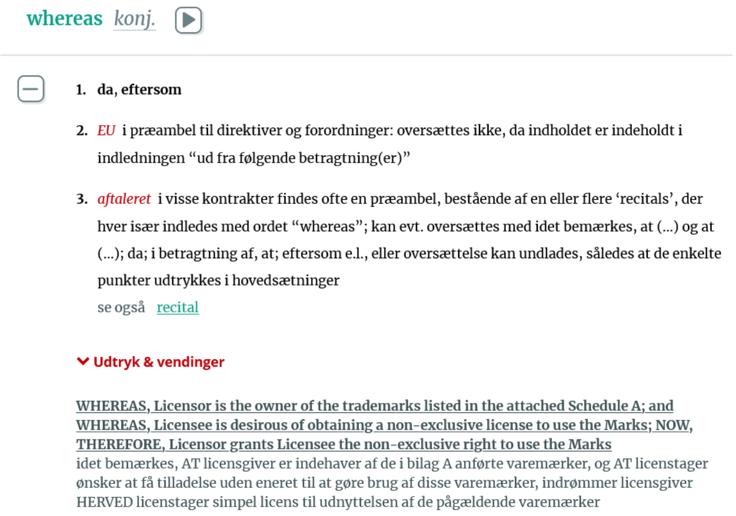
aksjekapital
Eng: share capital (but see comment below)
USA: stated capital
COMMENT In the USA the terms "capital stock", "capital shares" and "legal capital" are also used. The term "capital" is sometimes used both in England and the USA. But "capital" has many other meanings as well and might lead to confusion.
COMMENT ON TRANSLATION In England, once an issue of shares is fully paid, the company may convert the shares into "stock". Stock is capable of being divided into and held in any irregular fractional amount, while shares are of equal denomination. If a company has made such a conversion, the *aksjekapital** will be referred to as "stock capital".

Figure 7: Comment on translation in *Stor norsk–engelsk juridisk ordbok*

The explicit comment on translation in Figure 7 concerns the translation of the Norwegian term *aksjekapital* into British English (Eng) and combines a legal and a translation context. The lexicographer thus provides an extra-lexicographic

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1936 (Article)

context by explaining elements of substantive law (diatechnical dimension) and a use context by helping translators selecting the appropriate term. These contexts are aligned with the dictionary's user group in terms of factual and language competences when translating legal texts, i.e. user needs. An English–Danish law dictionary whose user group is legal translators, other translators, interpreters, lawyers, the business community, and students provides help to translation in very specific parts of the translation process (Figure 8).



whereas *konj.*

1. **da, eftersom**
2. *EU* i præambel til direktiver og forordninger: oversættes ikke, da indholdet er indeholdt i indledningen "ud fra følgende betragtning(er)"
3. *aftaleret* i visse kontrakter findes ofte en præambel, bestående af en eller flere 'recitals', der hver især indledes med ordet "whereas"; kan evt. oversættes med idet bemærkes, at (...) og at (...); da; i betragtning af, at; eftersom e.l., eller oversættelse kan undlades, således at de enkelte punkter udtrykkes i hovedsætninger
   se også recital

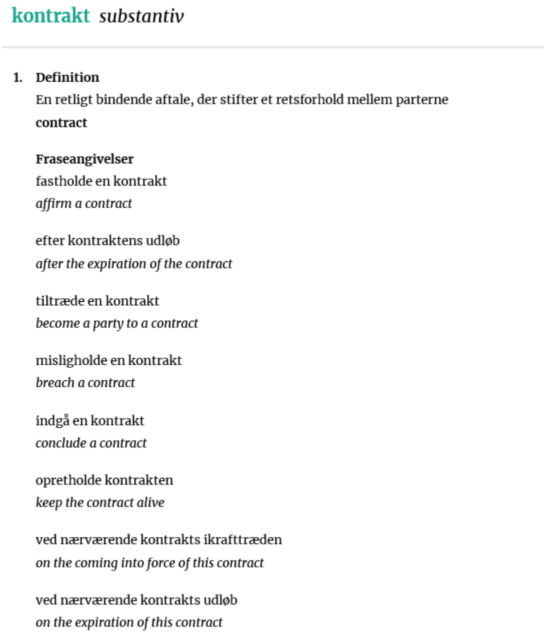
Udtryk & vendinger

WHEREAS, Licensor is the owner of the trademarks listed in the attached Schedule A; and WHEREAS, Licensee is desirous of obtaining a non-exclusive license to use the Marks; NOW, THEREFORE, Licensor grants Licensee the non-exclusive right to use the Marks
idet bemærkes, AT licensgiver er indehaver af de i bilag A anførte varemærker, og AT licenstager ønsker at få tilladelse uden eneret til at gøre brug af disse varemærker, indrømmer licensgiver HERVED licenstager simpel licens til udnyttelsen af de pågældende varemærker

Figure 8: Detailed help to translate English legal texts into Danish: *Juridisk ordbog engelsk–dansk*

The article in Figure 8 addresses context in several ways. Firstly, it concerns the conjunction *whereas* and presents three contexts: one unmarked, general translation option; one restricted to a European Union context (*EU*); and one restricted to contract law (*aftaleret*). Secondly, the data contained in the comment on EU translation explains that in the preamble of directives and regulations, the English conjunction *whereas* is not translated into Danish because the meaning of the conjunction is contained in the Danish introduction by the phrase "*ud fra følgende betragtning(er)*" (i.e. based on the following consideration(s)). This reference to the dictionary-external world indicates that the wording in specific text genres is structured differently in English and Danish within the same genre and the same supranational jurisdiction. Thirdly, the context provided for meaning 3 explains that English contracts often contain preambles with recitals introduced by the conjunction *whereas* and that the English conjunction can be translated in several ways, including non-translation if the Danish translator uses main clauses instead of conjunctions. Finally, the article contains an English example sentence from a contract and its Danish translation (*Udtryk & vendinger*), which explicitly shows how to translate the English conjunction *whereas* and the remainder example text into Danish by using main clauses. Example sentences like this illustrate usage in legal contexts at the level above words, terms, and phrases with a direct link to translation.

The example sentence in Figure 8 indicates that dictionaries can help users translate legal collocations and phrases. Alcaraz Varó and Hughes (2002: 167-170) argue that collocations are important parts of legal texts, and it is imperative that they are translated correctly in a legal sense as well as a language sense; legal collocations may broadly be described as lexical combinations regularly found in legal texts. Since collocations are rather short strings of text, lexicographers may consider including phrases in their dictionaries, for instance because they may be difficult to translate factually, grammatically, and idiomatically correct. This may be due to false friends, fixed collocations or phrases, different languages having different rules of syntax and morphology, all or some of which translators may be unaware of (see the discussion of user competences and skills in Section 2 above). An example of a Danish–English legal translation dictionary containing collocations and phrases is found in Figure 9.



kontrakt *substantiv*

1. **Definition**
En retligt bindende aftale, der stifter et retsforhold mellem parterne
contract

Fraseangivelser
fastholde en kontrakt
affirm a contract

efter kontraktens udløb
after the expiration of the contract

tiltræde en kontrakt
become a party to a contract

misligholde en kontrakt
breach a contract

indgå en kontrakt
conclude a contract

opretholde kontrakten
keep the contract alive

ved nærværende kontrakts ikrafttræden
on the coming into force of this contract

ved nærværende kontrakts udløb
on the expiration of this contract

Figure 9: Excerpt with collocations and phrases from *Juridisk ordbog engelsk–dansk/dansk–engelsk*

Figure 9 illustrates how a Danish–English legal translation dictionary intended for Danish translators, lawyers and students presents Danish word combinations and their translations into English. In addition to a definition in Danish and the boldface translation equivalent, the article contains a total of 14 Danish collocations with their English translations, such as *affirm a contract* and *conclude a contract*, and phrases, such as *on the coming into force of this contract* and *on the expiration of this contract*. The last two phrases show, for instance, that the Danish and English word order is different (compare the placing of the terms *kontrakt* and *contract*), and prevents the use of false friends, in that the Danish preposition *ved* generally corresponds to the English prepositions *with* and *by*, but the English preposition *on* has to be used when translating the legal phrases. This illustrates an important aspect of specialised translation, as explained by Gerzymisch-Arbogast (2008: 30): "It is the whole unit of such lexical expressions that the translator has to identify and transport to another textual environment." Therefore, translators will benefit from legal translation dictionaries containing such phrases.

Finally, it seems appropriate to note that contextual data in legal translation dictionaries should be accurate and sufficient. One challenge for lexicographers is that contextual data may be sufficient for one user group but not for another, depending on their legal, language, and translation competences and skills (see Section 2 above). If translators misunderstand the data and because of that produce translations containing incorrect legal terms, this may have serious financial and legal consequences. Contextual data referring to specifics of legal systems are often found in definitions or notes and, if inaccurate or insufficient, may lead to serious mistakes in translations. Legally incorrect translations may fail to create or transfer legal rights and obligations, which in turn may have severe financial and property-related consequences, for example, for parties to contracts and beneficiaries in wills. Such faulty translations may be caused by insufficient definitions or inaccurate collocations or phrases in dictionaries. In addition, references to dictionary-external sources may link to outdated information (see discussion of Figure 5 above) and thus mislead translators and affect the legal contents of translations negatively. Contextual data referring to specifics of legal language are often found in comments (see discussion of Figure 8 above) and should result in grammatically and idiomatically correct translations, but if insufficient may lead to ungrammatical or idiomatically incorrect translations. Such faulty translations may have none or insignificant legal and financial consequences but may result in translations that readers find have strange syntax and clause structures and therefore difficult to understand.

Having seen that contextual data are helpful in legal translation dictionaries, it is appropriate to examine how contextual data relate to the extra-lexicographic translation activity.

4. Context and the translation process

The inclusion of context in legal translation dictionaries may be explained further by examining the relation of context to two elements: the translation process

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1936 (Article)

and the language in which context is presented. A discussion of the translation process may thus take into consideration the extent to which dictionaries are intended to assist translators. A very general description of the translation process divides it into three phases: Translators decode (interpret and understand) the text to be translated, transfer the meaning of the text, and encode (produce) a text in another language (see e.g. Alcaraz Varó and Hughes 2002: 23; and Gerzymisch-Arbogast 2008: 12). From this perspective, legal translation dictionaries should contain data that enable translators to decode, transfer and encode, with or without the use of generative artificial intelligence, and this involves two languages. The article in Figure 9 contains a Danish term and a Danish definition providing semantic context, which may help translators to understand the meaning of the term in Danish source texts, i.e. an example of assistance in the decoding phase of translation. Figures 6 and 7 present domain-specific and translation-specific contexts to users and may thus be said to assist in the transfer phase of translation. The article shown in Figure 8 presents contextual data to users relating to the production of texts in such a way that it may be said to assist the encoding phase of translation. Finally, Figure 5 contains context expressed in Norwegian and German and Figure 6 contains context in Norwegian and English to Norwegian users, while Figure 9 provides context in Danish and English to Danish users, and these examples may be said to assist in all three translation phases. Finally, Figures 5, 6, and 7 contain comparative legal data assisting translators in their decision-making.

When discussing context in legal translation dictionaries and their relation to the translation process, lexicographers should consider the language in which they present context. To be as useful as possible, contextual data should be expressed in language that is easily understood by the dictionary's intended user group. The use of the native language of the users means that more will be able to understand the contextual data easily and correctly than if they had been written in a foreign language. Figures 2 and 7 appear to contradict this statement in that the language of contextual data is English despite the intended users being Norwegian. Nevertheless, the contextual data indirectly help users to translate Norwegian texts in a proper way, or otherwise communicate about Norwegian law. Figures 3, 4, 5, 6, 8 and 9 present contexts in the native language of the intended user groups. For instance, context is expressed in Danish in Figure 3, which contextualises French legal terminology, and context is partly expressed in Norwegian and partly in German in Figure 5. That being said, the contexts provided in Figures 2 and 7 could have been expressed in the native language of the intended user groups instead of English without loss of information, which may be substantiated by the Danish definitions in Figures 3 and 4, which cater for Danish user groups. The above considerations are subject to a proviso, however: The language, legal, cultural, and translational competences and skills of the intended user groups should be the determining factors when deciding on the language in which to present contexts.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1936 (Article)

5. Concluding remarks

This article set out to investigate context in legal translation dictionaries and how contextual data can help legal translators in the translation process. The findings indicate that context in relation to legal translation dictionaries can be described as a concept with two dimensions. The first is the overall context in which the dictionaries are intended to be used and relates to the extra-lexicographic translation situations and the intended user groups, which may be seen as a function-dependent dimension. The second dimension is the context(s) in which specific data types in the dictionaries are situated and relates to the selection and presentation of contextual data supporting translation of legal texts, which may be described as a data-dependent dimension. The above discussion reveals that the two dimensions interact with each other in that decisions lexicographers make in the data-dependent dimension rely on decisions made in the function-dependent dimension and vice versa. For example, online formats provide lexicographers with various digital options for finding and presenting contextual data to users and may offer users ways in which to access the data in legal translation dictionaries that differ from those offered by print dictionaries.

Law and legal language are culture-bound in the sense that they represent a system and a language tied to a particular jurisdiction such as a country or region. This affects the overall context in which legal translation dictionaries exist and their lexicographic scope in that they treat two or more legal systems and their associated languages. Lexicographers may choose between various solutions in their attempt to help translators such as diatopical and diatechnical labels, which help translators distinguish between jurisdictional term variants. At the same time, legal translation is an interdisciplinary activity involving competences and skills relating to law, language, and translation, meaning that legal translation dictionaries should provide help at various levels of translation such as terms, words, collocations, and phrases. Furthermore, legal translation involves a decoding, a transfer, and an encoding phase, each requiring different types of help from legal translation dictionaries. For dictionaries to provide the necessary help to legal translators, lexicographers should consider presenting legal concepts and terms with their definitions and comparative legal data, as well as examples of their occurrence. This may be done by showing pragmatic contexts, syntactic-semantic contexts, and contexts of use related to source-language as well as target-language items, including concepts, terms, collocations, phrases, translation equivalents, example sentences, dictionary-internal cross-references, and dictionary-external references. Dictionaries may include all the necessary context for a specific translation task, but lexicographers are not responsible for what dictionary users do with the dictionary data. No matter how much context legal translation dictionaries provide, there is no guarantee that translations will be correct.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1936 (Article)

References

A. Dictionaries

Craig, R.L. 1999. *Stor norsk–engelsk juridisk ordbok. Med engelsk–norsk register* [Norwegian–English Law Dictionary. With English–Norwegian Index]. Oslo: Universitetsforlaget.

Dear, I.C.B. and P. Kemp (Eds.). 2007. *The Oxford Companion to Ships and the Sea*. Second edition. Oxford: Oxford University Press.
DOI: 10.1093/acref/9780199205684.001.0001

Frandsen, H.P. (n.d.). *Juridisk ordbog engelsk–dansk* [Law Dictionary English–Danish]. Odense: Ordbogen.com.

Hartmann, R.R.K. and G. James. 2001. *Dictionary of Lexicography*. Revised edition. London/New York: Routledge. Available at:
https://doi.org/10.4324/9780203017685

Hjelmblink, S. 1991. *Retsplejeordbog. Dictionnaire Juridiciaire. Judicial Dictionary. Prozesswörterbuch.* Copenhagen: Munksgaard.

Lind, Å. 2003. *Norsk–engelsk juridisk ordbok. Sivil- og strafferett* [Norwegian–English Dictionary of Law. Civil and Criminal Law]. Oslo: Cappelen.

Nielsen, S. 2023. *Juridisk ordbog engelsk–dansk/dansk–engelsk* [Law Dictionary English–Danish/Danish–English]. Odense: Ordbogen.com.

Simonnæs, I. 1994. *Norsk–Tysk Juridisk Ordbok* [Norwegian–German Dictionary of Law]. Bergen: Fagbokforlaget.

B. Other literature

Alcaraz Varó, E. and B. Hughes. 2002. *Legal Translation Explained*. Manchester: St. Jerome. Available at: https://doi.org/10.4324/9781315760346

Baclawski, K., M. Bennett, G. Berg-Cross, C. Casanave, D. Fritzsche, J. Luciano, T. Schneider, R. Sharma, J. Singer, J. Sowa, R.D. Sriram, A. Westerinen and D. Whitten. 2018. Ontology Summit 2018 Communiqué: Contexts in Context. *Applied Ontology* 13(3): 181-200. Available at: https://doi.org/10.3233/AO-180200

Bajčić, M. 2017. *New Insights into the Semantics of Legal Concepts and the Legal Dictionary*. Amsterdam: John Benjamins. Available at: https://doi.org/10.1075/tlrp.17

Bergenholtz, H. and U. Kaufmann. 1997. Terminography and Lexicography. A Critical Survey of Dictionaries from a Single Specialised Field. *Hermes. Journal of Linguistics* 18: 91-125.

Bergman, M.K. 2018. *A Knowledge Representation Practionary. Guidelines Based on Charles Sanders Peirce*. Cham: Springer. Available at: https://doi.org/10.1007/978-3-319-98092-8

Biel, Ł. 2022. Translating Legal Texts. Malmkjær, K. (Ed.). 2022. *The Cambridge Handbook of Translation*: 379-400. Cambridge: Cambridge University Press. Available at:
https://doi.org/10.1017/9781108616119.020

Council of the European Union. 2022. Convention on the Recognition and Enforcement of Foreign Judgments in Civil or Commercial Matters. *OJ L* 187, 14.7.2022: 4-15. Available at:
https://eur-lex.europa.eu/legal-content/EN-DA/TXT/?from=EN&uri=CELEX%3A22022A0714%2801%29

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1936 (Article)

De Groot, G.-R. 1990. Die relative Äquivalenz juristischer Begriffe und deren Folge für mehrsprachige juristische Wörterbücher. Thelen, M. and B. Lewandowska-Tomaszczyk (Eds.). 1990. *Translation and Meaning: Proceedings of the Maastricht Session of the 1990 Maastricht-Lodz Duo Colloquium on "Translation and Meaning," Held in Maastricht, The Netherlands, 4–6 January 1990.* 122-128. Maastricht: Euroterm. Available at: https://cris.maastrichtuniversity.nl/ws/portalfiles/portal/46991728/d1937179-8a52-49a4-afbc-e0781cace612.pdf

Domínguez Vázquez, M.J. and R.H. Gouws. 2023. The Definition, Presentation and Automatic Generation of Contextual Data in Lexicography. *International Journal of Lexicography* 36: 233-259. Available at: https://doi.org/10.1093/ijl/ecac020

Fluck, H.-R. 1985. *Fachsprachen. Einführung und Bibliographie.* Third, updated and expanded edition. Tübingen: Francke.

Fuertes-Olivera, P.A. (Ed.). 2018 *The Routledge Handbook of Lexicography*. Abingdon/New York: Routledge. Available at: https://doi/10.4324/9781315104942

Gerzymisch-Arbogast, H. 2008. Fundamentals of LSP Translation. *MuTra* 2: 7-64. Available at: http://www.translationconcepts.org/pdf/MuTra_Journal2_2008.pdf

Kipfer, B.A. 2022. Glossary of Lexicographic Terms. Jackson, H. (Ed.). 2022. *The Bloomsbury Handbook of Lexicography.* Second edition: 441-457. London: Bloomsbury Academic.

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. Available at: http://wa.amu.edu.pl/~rlew/pub/Lew_space_restrictions_in_paper_and_electronic_dictionaries.pdf [1 March 2024].

Nielsen, S. 1990. Contrastive Description of Dictionaries Covering LSP Communication. *Fachsprache/International Journal of LSP* 3–4: 129-136.

Nielsen, S. 2018. Lexicography and Interdisciplinarity. Fuertes-Olivera, P.A. (Ed.). 2018. *The Routledge Handbook of Lexicography*: 93-104. Abingdon/New York: Routledge.

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. Available at: https://doi.org/10.1075/hot.3.leg4

***Oxford Reference Online*.** https://www.oxfordreference.com/browse [20 January 2024].

Prieto Ramos, F. 2021. Translating Legal Terminology and Phraseology: Between Inter-systemic Incongruity and Multilingual Harmonization. *Perspectives* 29(2): 175-183. Available at: https://doi.org/10.1080/0907676X.2021.1849940

Sandrini, P. 2018. Legal Translation. Humbley, J., G. Budin and C. Laurén (Eds.). 2018. *Languages for Special Purposes: An International Handbook*: 548-562. Berlin/Boston: De Gruyter Mouton. Available at: https://doi.org/10.1515/9783110228014-029

Tarp, S. 2018. The Concept of Dictionary. Fuertes-Olivera, P.A. (Ed.). 2018. *The Routledge Handbook of Lexicography*: 237-249. Abingdon/New York: Routledge.

Zgusta, L. 1971. *Manual of Lexicography*. [in cooperation with V. Cerny, Z. Hermanovä-Novotnä, D. Heroldovä, L. Hrebicek, J. Kalouskovä, V. Miltner, Y. Minn Latt, L. Motalova, K. Peträcek, K.F. Ruzioka, I. Vasiljev, P. Zima and K. Zvelebil]. Prague/The Hague: Academia/Mouton. Available at: https://doi.org/10.1515/9783111349183

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1937 (Article)

Making African Dictionaries More African

Thapelo J. Otlogetswe, *Faculty of Humanities,*
University of Botswana, Gaborone, Botswana
(otlogets@ub.ac.bw) (https://orcid.org/0000-0003-3887-2982)

Abstract: This article proposes strategies for compiling African dictionaries that are Afrocentric. It argues that such dictionaries must deal appropriately with the complex morphology typical of many African languages. Addressing complex morphology will help users connect words and meanings to develop a more nuanced understanding of the language. Afrocentric dictionaries ought to provide historical and etymological information of their entries, tracing lexical origins, migrations, and influences which would enrich the understanding of language evolution and its interconnectedness. For a broader vocabulary coverage related to African culture, traditions, flora, fauna, geography, and history, the article proposes that African dictionaries need to adopt a hybrid strategy of data collection that engages linguists, scholars, native speakers and community members in the dictionary-making process. Additionally, the paper argues that the middle section of African dictionaries presents an excellent space for lexicographers to capture, demonstrate, and preserve African cultures. The section should affirm, reinforce, and celebrate African food, attire, dances, cattle colour terminology — especially for pastoral communities — and kinship terminology. By implementing approaches set out in this article, it is hoped that African dictionaries will become more representative, inclusive, and reflective of the diverse African languages, cultures, and contexts.

Keywords: AFROCENTRIC, SETSWANA, AFRICAN DICTIONARIES, ETYMOLOGY, CULTURE, MORPHOLOGY

Opsomming: Om Afrika-woordeboeke meer Afrosentries te maak. In hierdie artikel word Afrosentriese strategieë vir Afrika-woordeboeke voorgestel. Daar word aangevoer dat hierdie woordeboeke die komplekse morfologie, wat tipies van Afrikatale is, op 'n gepaste wyse moet hanteer. Die aanspreek van die komplekse morfologie sal gebruikers help om woorde en betekenisse te verbind om sodoende 'n meer genuanseerde begrip van die taal te ontwikkel. Afrosentriese woordeboeke behoort historiese en etimologiese inligting van hul inskrywings te verskaf deur die leksikale oorsprong, migrasies, en invloede na te spoor, wat die begrip van taalevolusie en die onderlinge samehang daarvan sal verryk. Vir 'n breër woordeskatdekking wat verwant is aan die Afrika-kultuur, -tradisies, -flora, -fauna, -geografie en -geskiedenis, word daar in die artikel voorgestel dat Afrika-woordeboeke 'n hibriede strategie van data-insameling moet volg wat taalkundiges, vakkundiges, moedertaalsprekers en gemeenskapslede in die woordeboekmaakproses betrek. In hierdie artikel word ook geargumenteer dat die middelste deel van Afrika-woordeboeke 'n uitstekende spasie bied vir leksikograwe om die Afrika-kulture vas te vang, te demonstreer en te bewaar. In hierdie afdeling moet Afrika-kos, -kleredrag, -danse, -veekleurterme — veral vir landelike gemeenskappe — en verwantskapsterme verklaar, versterk, en gevier word. Deur benaderings, soos in hierdie artikel uiteengesit, te implementeer, word daar gehoop dat Afrika-woordeboeke

Lexikos 34 (AFRILEX-reeks/series 34: 2024): 309-330

meer verteenwoordigend, inklusief, en weerspieëlend van die diverse Afrikatale, -kulture, en -kontekste sal word.

Sleutelwoorde: AFROSENTRIES, SETSWANA, AFRIKA-WOORDEBOEKE, ETIMOLOGIE, KULTUUR, MORFOLOGIE

1. Introduction

Africa resonates with a rich tapestry of languages and cultures. Such a scenario presents unique challenges and opportunities for lexicographers working on the continent to craft dictionaries that are not mere linguistic compendiums, but living embodiments of African heritage, identity, and empowerment. With its staggering linguistic diversity, Africa is a continent of untold stories, intricate narratives, and boundless creativity. Yet, we find that dictionaries that were framed through lenses foreign to the continent's experiences, fail to capture the depth, and nuances that define the myriad tongues. The challenge that faces African lexicographers is to explore methodologies that pay homage to the authenticity of African languages. They must seek to forge a path towards dictionaries that encapsulate the very essence of Africa's diverse cultures. African lexicographers need to unravel the threads that tie language to identity, understanding that a dictionary is not merely a lexical inventory, but a living testament to the collective memory of a people. It is a repository of wisdom, a mirror to shared experiences, and a beacon for future generations. This article therefore argues that African lexicographers ought to embark on a collective mission to celebrate and preserve the linguistic heritage of the continent, fostering a sense of belonging and pride in every word, phrase, and expression that graces the pages of an African dictionary.

At the heart of such an enterprise, lies the question of how to make African dictionaries, specifically general dictionaries, truly or uniquely African. Features proposed in this article could be used as a checklist, or as tools that can be used to measure how African or reflective of a linguistic community a dictionary is. Historically, most African dictionaries came to communities through English or French missionaries (Pawliková-Vilhanová 2009, Prah 2009). These were mostly bilingual dictionaries serving a translation function in missionary schools or aiding evangelical work. Prinsloo and Zondi (2020: 17) refer to such dictionaries when they observe that:

> The history of lexicography in Africa began because of European activities: exploration, evangelization, and colonialization. The early lexicons, whether compiled by explorers, missionaries, or colonial administrators, were "Euro-centred", produced in Europe for Europeans rather than for African users. … Even with the emergence of modern linguistics, lexicographic works have been primarily intended for scholarly interest *and not for the needs of ordinary Africans* (my emphasis).

While this may be the context within which earlier African dictionaries were compiled, African dictionaries, do not have to mirror European dictionaries, since

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1937 (Article)

African dictionaries exist in a unique environment which may necessitate a different structure and a different data collection strategy from those in Europe or America. What is needed are dictionaries designed to serve and reflect African communities.

The concern over African dictionaries with a Eurocentric bias is not new. Gangla (2001: 52) has decried Eurocentrism in African lexicography when he demonstrated that the pictorial illustrations of the *South African Multi-languages Dictionary and Phrase Book* (Reynierse 1996) were more Eurocentric than African. This, he argued, was shown by illustrated elements that did not reflect typical African architecture and sports. The dictionary instead depicted rooms in a house with a pantry, and sports played, being cricket and rugby.

Prinsloo (2017a: 5) has gone further to attempt a characterization of African dictionaries based on five classifications:

i. Dictionary compilations by foreigners abroad. He classified these as having a "true Euro-centric approach."
ii. Dictionary compilations by foreigners in Africa, e.g., on missionary posts using Africans as informants. He classified these as having a "Euro-centric approach."
iii. Dictionary compilations by non-mother-tongue speakers of African languages who studied the grammar and even learned to speak African languages, working with mother-tongue speakers. He classified these as "containing Afro-centric elements."
iv. Dictionary compilation by Africans guided by foreigners. He classified these as "containing Afro-centric elements."
v. Dictionary compilation by Africans. He classified these as having a true Afro-centric approach.

Prinsloo's classification is an important contribution to determining a dictionary's Afrocentricity though it focuses much on the identity of a compiler and not on what is definitive of an Afrocentric or Eurocentric dictionary. There needs to be more thought into crafting features or measures of dictionary Afrocentricity. Such features can inform compilers at a dictionary design stage or function as a yardstick through which a dictionary's Afrocentricity could be measured. With the dictionary tradition in Africa having been established by non-Africans, it is conceivable to have Eurocentric dictionaries compiled by Africans. What practical lexicographers need is a clearer picture of what the features of an Afrocentric dictionary are.

2. Characteristic features of an Afrocentric dictionary

This article proposes six features that should characterise any general dictionary in Africa. These features are the use of multiple word-collection strategies, dialectal variation, inclusion of non-standard and colloquial words, the treatment of the unique morphology of a language, etymology inclusion, and a compre-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1937 (Article)

hensive coverage of the culture of a speech community of the dictionary.

2.1 Use multiple word-collection strategies

The unique African context demands that African dictionaries should be compiled through multiple-word collection strategies. This is because most African languages have a poor literary tradition and cannot depend exclusively on corpora. For many African languages with some written materials, such texts are usually within the restricted domain of creative works. Most African languages are not used in science, technology, engineering, and law domains. Some African languages are so underdeveloped that they lack standard orthographies, and their grammars are still undeveloped. This means that while the field of lexicography internationally has experienced an explosion of corpus lexicography (cf. Granger and Paquot 2012, Hanks 2012), many African language dictionaries cannot be compiled exclusively on corpora evidence (cf. Prinsloo 2017b). This is not to say that corpora are not important. They are indeed important. Atkins et al. (1992: 10) have argued that "[i]n our ten years' experience of analysing corpus material for lexicographical purposes, we have found any corpus — however 'unbalanced' — to be a source of information and indeed inspiration. Knowing that your corpus is unbalanced is what counts." The situation as defined by Atkins et al. is typical of very few African languages which have corpora. The vast majority of African languages lack corpora and should consider alternative methods of data collection for dictionary compilation. It is no wonder that Nkabinde (2003: 174) argues that "[a] corpus should be used to supplement the usages obtained through fieldwork in the compilation of a dictionary."

The major languages of the world such as English, French, Spanish and Portuguese often have large, published dictionaries available to them. They are usually supported by major publishing companies, such as Oxford University Press, Cambridge University Press, Longman, and Macmillan Publishers, who employ scores of professional lexicographers to compile massive text corpora and do the necessary research to compile quality dictionaries (Barati and Noor 2011, Robbins 2017). For many African languages, as Moe (2003: 216) has demonstrated, the picture is far bleaker. With few or no published texts, few, or no professional lexicographers available to them, and little or no funding, many African languages face a daunting challenge.

The challenge is even greater for many Khoisan languages such as N|uu who are on the verge of extinction (Sands and Jones 2022). Prinsloo (2015) also demonstrates that the dictionary compilation of Ju|'hoasi was compiled by a partnership between native-speaker children and academics. This is true for the compilation of the *N|uuki Namagowab Afrikaans English dictionary* (Sands and Jones 2022). Such collaborative efforts foster accuracy, inclusivity, and cultural relevance. African lexicographers should therefore use multiple word collection strategies so that the final dictionary captures regional and local variations of a language.

A hybrid strategy of data collection in African dictionaries compilation that engages linguists, scholars, native speakers, and community members in the dictionary-making process is attractive since it ensures a broader coverage of the lexicon. Having a broader team of contributors contributes to a broader headword list that covers a wide range of terms relating to various elements of African culture. These include local expressions, proverbs, idioms, and other cultural nuances that enrich a dictionary. The data collection strategy must also include consulting other collections of knowledge compilations in the language such as collections of herpetology (Auerbach 1986), animals and plants (Cole 1995) to enrich a dictionary with elements of indigenous knowledge systems. This means that African lexicographers ought to read widely in the related fields of culture, history, and anthropology to benefit from wide scholarship on the language, and culture of the linguistic community. For example, a study in Tswana hunting practices in the Transvaal and Kalahari (Morton and Hitchcock 2014) has contributed a rich hunting vocabulary that has enriched the second edition of *Tlhalosi ya Medi ya Setswana* which is still under development. Such vocabulary includes the following Setswana words:

- **mosokela-tsebeng** bow string
- **mutlwana**(e)i snare
- **segole** whip snare
- **telekelo** a hunting pit; a place to which animals are chased
- **tlhabadilebanye** bowstring
- **tlhagare** small iron arrow tip
- **tlhobolo** quiver for arrows (archaic; adopted for firearm, gun, rifle)
- **tshane** broad-bladed spear; sharpened stick used by herdboys
- **tshosa** long spear with a large blade

Finally, Moe (2003) has proposed a word collection strategy that uses a complex system of over 1800 semantic domains which has been proven to speed up word collection by yielding about 10,000 words per work. The semantic domains are now part of Fieldworks Language Explorer (Flex) which is used in dictionary compilation in Africa (Beier and Michael 2022). This method of word collection is one that lexicographers working on African dictionaries could explore to widen their data capture.

2.2 Capture dialectal variation

The second essential element for African lexicographers to consider is the representation of all dialects of their language since most African languages are not homogenous but are characterised by significant variation. Most of the linguistic variations are according to the geographical distribution of speakers who display differences in terms of vocabulary, grammar, and pronunciation (Southerland and Katamba 1996: 565, Honey 1997: 3, Crystal and Davy 1969: 67). A general dictionary of an African language should reflect dialectal varieties of a language. It must not restrict itself to one or a few dialect(s) that a compiler judges as correct or pure. African lexicographers should move away from the narrow representation of one dialect in a general dictionary. They must capture all the dialects, tag them, and cross-reference the synonymous entries, especially words from regional speakers of a language. Lexical differences play a significant role in regional dialectology and inclusive lexicography since it reflects the lifestyles that diverse groups live or have lived (Batibo 1999). For instance, Otlogetswe (2012) has entered and cross-referenced all four Setswana dialectal words that refer to pounded meat. These are *seswaa*, ***tšhôtlhô***, *tshwaiwa*, and *loswao*.

— **seswaa** /sìswàá/ Ngwt. •*ln. 7. se-, *8. di-*• nama e e budusitsweng thata e bo e kgobiwa *Nna o ntsholele seswaa rra, meno a me ga a thata* = TŠHÔTLHÔ, LOSWAO, TSHWAIWA
— **tšhôtlhô** /ʧʰɔ̀tɬʰɔ̀/ Kgat. •*ln. 9. n-, *10. din-*• nama e e swailweng kgotsa e thugilwe e budule = SESWAA, LOSWAO, TSHWAIWA
— **tshwaiwa** /ʦʰwàíwà/ Kwen. •*ln. 9. n-*• = LOSWAO, SESWAA, TŠHÔTLHÔ,
— **loswao** /lʊ̀swàʊ̀/ Ngwk. •*ln. 11. lo-*• **1**◉ nama e e budusitsweng thata e bo e kgobiwa = TŠHÔTLHÔ, SESWAA, TSHWAIWA **2**◉ Ngwt. thobane e e dirisiwang go swaa nama.

The four entries are tagged for the Setswana dialects of Sengwato (Ngwt.), Sekgatla (Kgat.), Sekwena (Kwen.) and Sengwaketse (Ngwk.).

2.3 Include non-formal, non-standard, and colloquial words

Lexicographers, in general, agree that a dictionary should document all frequently used words and meanings used in a language (Gouws and Prinsloo 2005, Svensén 2009). They agree that a compiler must not be prescriptive and attempt to choose the words he or she judges as good in the language since such a biased capture of the language will misrepresent the language and mislead the user. The lexicographer's role is to collect "all the words, whether good or bad" (Trench 1857: 7). This means that non-formal, non-standard, and colloquial words need to be collected and lemmatised. A lexicographer should primarily see himself as a historian and not a language critic. He or she must not see the dictionary as a representation of "the standard language," but rather as a collection of words used by speakers of the language, whether standard or not (Otlogetswe 2013). This argument is not new. Trench (1857: 7) has argued before that:

> A dictionary … is an inventory of the language … It is no task of the maker of it to select the *good* words of a language. If he fancies that it is so, and begins to pick and choose, to leave this and to take that, he will at once go astray. The business which he has undertaken is to collect and arrange all the words, whether good or bad, whether they do or do not commend themselves to his judgment, which, with certain exceptions hereafter to be specified, those writing in the lan-

> guage have employed. He is an historian of it not a critic … There is a constant confusion here in men's minds. There are many who conceive of a Dictionary as though it had this function, to be a standard of the language; and the pretensions to be this which the *French Dictionary of the Academy* sets up, may have helped this confusion. It is nothing of the kind.

A general African dictionary should therefore be inclusive of all speech registers of a linguistic community and not discriminate against others. It must mark informal, colloquial, and offensive entries but not exclude them from a dictionary.

2.4 Respond to the unique morphological treatment of the language

African lexicographers need to walk away from an all-size-fits-all approach to dictionary compilation, and instead investigate features of their languages that demand special treatment, to shape a more definite language-specific macrostructure and microstructure. Specific attention needs to be paid to the morphology of a language and how it can be best represented in a dictionary.

The morphology of many African languages is complex, and it is essential that lexicographers respond to the morphological peculiarities of the languages they are compiling (Prinsloo 2015). Gouws and Prinsloo (2005) have demonstrated that dealing with complex morphology can help users make connections between words and meanings. It can help learners and dictionary users develop a more nuanced understanding of the language. Many lexicographers have rightly critiqued stem-based dictionaries and instead favoured word-based lemmatization (cf. e.g. De Schryver 2010, Prinsloo 2009, Van Wyk 1995). In most cases, they have done so without retaining the benefits that were addressed by the stem-based dictionaries and have not found an elegant way of linking semantically and morphologically related entries spread across a dictionary because of alphabetization (Van Wyk 1995).

A case in point is the marking and handling of deverbatives in Setswana dictionaries. An improved analysis of the morphology of deverbatives can reveal philosophical and conceptual views of the world hidden in the Setswana lexicon. These include nouns such as *boipuso* "independence, self-rule" derived from *busa* "govern, rule;" *boitumelo* "happiness" derived from *dumela* "believe, trust" and *kagiso* "peace" from *aga* "build."

The Setswana word *kagiso* "peace" is derived from the root verb *aga* "to build, to construct or to put together." The causative suffix *–is-* is then attached to the verb *aga* to form *agisa* "help someone build." The noun maker suffix [-o] is then attached to "agisa" and [k] inserted at the beginning to derive the noun *kagiso*. The complete process appears as follows:

aga + -is- = agisa + -o = agiso > [k] + agiso = kagiso

The central argument here is that instead of simply giving the meaning of a word, such as *kagiso* as "peace," a dictionary must fully capture the morphology

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1937 (Article)

of an entry to reveal its source to contribute to the understanding of the worldview of a linguistic group. For instance, an analysis of *kagiso* demonstrates that Batswana believe that *peace* is something that is constructed as shown by its roots from *aga* "build." It is seen as a progressive matter and not just a state.

Another example is of the word *boitumelo* that means "happiness". It is derived from the verb *dumela* "believe or accept." Analysed this way, Batswana can make inferences that in Setswana happiness is attained when one reaches a state of self-belief, a state of believing themselves. *Boitumelo* can therefore be decomposed into bo-itum-el-o.

African dictionaries must also make morphological links between verbs that are derived from other base verbs as in Table 1.

Table 1: Verbs derived from other base verbs

Verb	**English translation**	**Derived from**	**English translation**
Ipusolosetsa	to revenge	*busa*	return
Leseletsa	carry a bucket on top of the head without touching it	*lesa*	leave, do not touch
Itshephisa	make holy	*tshepha*	trust
Ikemisetsa	intend/purpose	*ema*	stand
Ikgolaganya	connect yourself with	*golega*	tie/bind
Ikgoga	drag yourself	*goga*	pull/drag

African lexicographers should therefore be sensitive to the morphology of the languages they work in and attempt to capture its richness.

2.5 African dictionaries and etymology

Most African dictionaries are of languages that are in contact situations where speakers speak or are exposed to multiple local or foreign languages. It is critical that African dictionaries provide historical and etymological information of words, tracing their origins, migrations, and influences. This enriches the understanding of language evolution and its interconnectedness. A better documented contact situation in scholarship is the one in which an African language has borrowed words from Afrikaans, French, English, or Portuguese (Matiki and Ramaeba 2018). Such borrowed terms usually cluster around certain semantic clusters such as food, education, farming, computing, government etc. For instance, in the general domain of food, maize has its roots in the Americas and came into Africa by Portuguese traders in the 1500s to the 1600s through Mozambique into South Africa (Miracle 1965, McCann 2001). Miracle demonstrates that linguistic evidence demonstrates that maize is not indigenous to Africa. This is reflected in the name for maize in many African languages. In the Semitic language spoken in Ethiopia maize is called *yabaher mashela* which means "grain from the sea." In Chichewa, maize is *chimanga* which means "from the coast." In Swahili maize is *muhindi* which means "a grain from India." Speakers of Kikongo call maize *maza mamputo* which means "white people's grain." In Mande, maize is *tuba-nyo* which means "a white people's grain." In Setswana, maize is *mmidi,* a word which came into Setswana sometime after 1700 from Afrikaans *mielie*, which itself is a borrowing from Portuguese *milho*. A single grain of maize in Setswana is called *pidi,* from Afrikaans *pit*. From the Afrikaans *kaboemielies,* Setswana has derived the word *kabu* which refers to boiled maize grains. Such linguistic evidence derived from history and archaeology can assist in enriching the microstructure of an African dictionary.

An uncommon and yet an important link in African dictionaries is to show borrowing between African languages themselves. For instance, Gunnink (2020) has demonstrated that Setswana has borrowed some of its vocabulary from Khoisan languages such as Khwe and Gǀui. Such words include *mokgalo* /mʊ̀qʰálɔ́/ "buffalo thorn (*Ziziphus mucronata*)" which is from Gǀui ǂqχ'árò, Kua ǂq'árò, or Naro ǂx'árò and *mokgwa* /mʊ̀qʰʷà/ "*Acacia nigrescens*" which is derived from Gǀui ǁqχ'úà or Kua ǁχ'úà. Setswana has also borrowed from Zulu and other African languages such as Kalanga (Otlogetswe 2016). For instance, Otlogetswe (2012) demonstrates that the Setswana word *tandabala* "old age pension for Botswana citizens who are 65 years and older" is borrowed from Kalanga. In Kalanga *tandabala* means "to sit with stretched out legs."

Another way in which African languages dictionaries could be enriched with etymology information is for them to include proto-Bantu roots to demonstrate links between African languages themselves (Batibo 1996). For instance, Table 2 demonstrates how Setswana entries could be associated with Proto-Bantu roots. If this is done across multiple Bantu languages, extremely useful links could be established between multiple languages.

Table 2: Proto-Bantu and Setswana terms

Proto-Bantu	**Setswana**	**gloss**
*-bab-	*baba*	itch, sting, hurt
*-gwina	*kwena*	crocodile
*-bac-	*batla*	to look for
*-bad-	*bala*	to count
*-pala	*phala*	impala
*-bada	*sebala*	spot, speckle
*-badi	*bedi; pedi*	two
*-beede	*lebele*	breast; udder
*-bin-	*bina*	dance
*-bid	*bela*	to be cooked; to boil
*-bod	*bola*	to be rotten
*-bombo	*mmopo*	forehead; bridge of nose; nose

Accounting for the etymology of African languages must be more than tracing borrowed words to their European heritage. It must also account for words between African languages themselves.

2.6 Capture the culture of speakers in detail

A general African dictionary must capture the rich culture of its linguistic group. Its pages must ooze with the prototypical elements of that culture. Culture refers to the shared beliefs, customs, behaviours, and artifacts that characterize a particular group or society. It was defined in 1871 by Edward B. Tylor as "that complex whole which includes knowledge, beliefs, art, morals, law, customs and any other capabilities and habits acquired by man as a member of society" (Tylor 1871: 1). Peterson (1979) argues that culture in contemporary parlance consists of four elements: norms, values, beliefs, and expressive symbols. Scholars have suggested various elements that are definitive of culture, including language, religion, cuisine, social habits, music, art, and literature (Spencer-Oatey 2012). A dictionary that is Afrocentric should therefore reflect an African culture in its macrostructure. Sociologists (Spencer-Oatey 2012, Peterson 1979) have proposed norms, language, festivals, rituals and ceremonies, holidays, pastimes, food, architecture, religions, values, taboos, sports, clothing and outfits, music, social hierarchy, symbols, dance, and art as cultural elements. These provide some guidance on what constitutes a more authentic culture of a people. Compilers of African dictionaries must consider the definitive cultural elements of the linguistic community within which they work. Nkabinde (2003: 170) has argued that "[o]ne of the methods of identifying words in cultural context is to determine the various cultural aspects peculiar to a speech community rather than address the entire domain of culture as such." He (ibid.) identified the following key features of Zulu culture: "kinship and social organization, communal activities, recreation, customs and beliefs, food and beverages, sorcery and witchcraft, mourning, burial and death, agriculture and animal husbandry, mythology and legends, music, dancing and song, folklore, national ceremonies." Such features cannot be assumed to be general to every African linguistic group. It is up to the compiler to identify cultural elements that are definitive of the linguistic community whose dictionary they are compiling and reflect it in the dictionary. They should seek to protect the culture and not undermine it — whether intentionally or subconsciously. They must affirm that African communities have their own food, attire, dances, unique ways of measuring time, cattle colour terminology — especially for pastoral communities, kinship terminology, etc.

We demonstrate this by exploring cattle and beef terminology of Batswana who are a pastoral community. Like many Bantu speakers, Batswana have a strong relationship with cattle reflected by a rich cattle terminology. They have a long tradition of cattle butchering that is deeply rooted in their cultural activities and rituals. The rituals are characterised by several meat dishes which are interwoven with their wedding, death, and birth rituals. Their vocabulary is rich of colour terms, lexicalized patterns, and colour placement on cattle skins. All these can inform and enrich the middle section of a general Setswana dictionary in a unique manner.

For instance, Figure 1 could be included in a Setswana general dictionary's middle section to show how Batswana partition a beef carcass in a distinct manner.

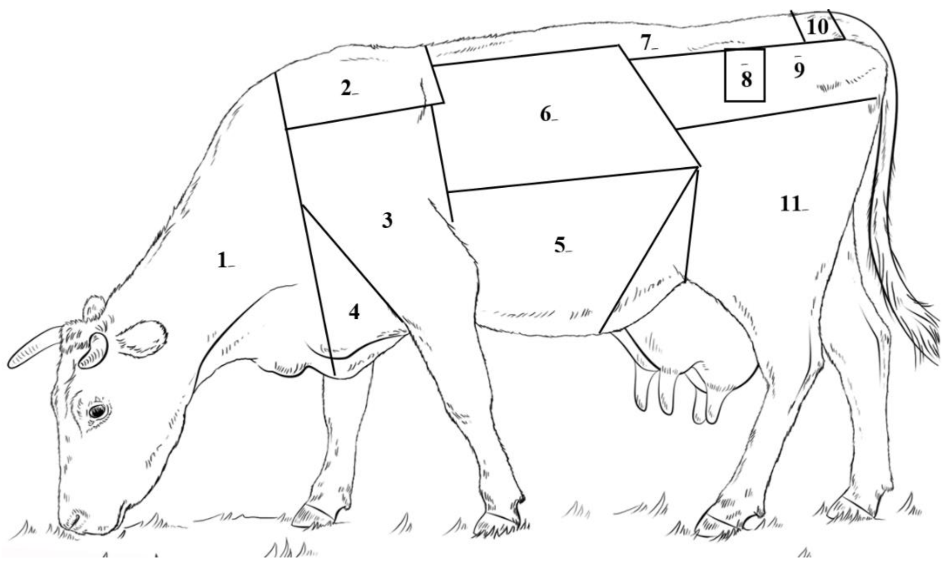


Figure 1: Setswana beef cuts (Adapted from Otlogetswe 2020)

Such an image could be accompanied by culturally relevant information explaining the different cuts as follows:

No 1. *Tlhogo* "head" which comprises the neck [*thamo*] and the first four ribs from the neck (two on the right and two on the left) (Otlogetswe 2019). *Tlhogo* is wrapped in a skin [letlalo] and given to the groom's maternal uncle who heads the wedding negotiations on the groom's side. He is the principal figure in a Setswana wedding and the ceremonies may not take place without his presence. As head of the wedding, he gets the *tlhogo*, the head of the slaughtered beast.

No 2. *Mokuana* (pl. *mekuana*): This is the meat found from just under the armpit of both front legs to above the shoulder of the cow. The shoulder bone is removed to allow the meat from the shoulder bone's cartilage to the spine to be cut out. This is done on both sides of the cow. During a cultural ceremony if a cow was shot, *mokuana* is given as payment to the person who shot the cow; if it was slaughtered, it is shared amongst the people who assisted.

No. 3. *Letsogo* (pl. *matsogo*). The word *letsogo* means an arm. However, in a cow it is the front leg, which is one of the first parts from a cow to be removed after it has been skinned. However, before both *matsogo* are removed, the metatarsal and metacarpal bones [*metwane*] are removed and thrown into a pot with miscellaneous pieces of meat, which are eaten after skinning. This pot of miscellaneous pieces is called *mothobiso*. *Matsogo* are dislodged from a cow with great precision. The lymph nodes [*dikgeleswa*] found where *matsogo* attach to the body, at the neck, must remain attached to the legs. One complete *letsogo* is given to the traditional doctor [*ngaka ya Setswana*] who ministered to the ceremony for his services. His role is to strengthen the wedding, to protect the homestead where the wedding is held, to doctor the fireplace, and to protect the bridal couple and all attendees from witchcraft. He is sometimes known as *setimamolelo* "the fire extinguisher" because the effects of witchcraft are usually expressed as heat.

No. 4. *Bokwana*. This is a cow's tip of the chest, which comprises fatty meat and bone. It is cooked and eaten by married individuals at the bride's home on the day that the bride is brought to the groom's homestead, the day after the wedding celebrations. The newlywed couple is served one large piece of meat on a plate for them to share. When served, this piece of *bokwana* is not cut with a knife since this would be considered a bad omen portending conflict between the couple. The couple eats by biting off pieces and placing the remainder back on the plate until it is finished.

No. 5. S*ehuba*. This is the chest area of a cow, which usually has a lot of fat, and includes all ribs that remain after the removal of *dithupa*. During the wedding, it is eaten by married men and women. It is believed that by eating it, partners affirm that they belong to each other via their chest, breasts, and heart.

No. 6. *Thupa* (pl. *dithupa*). The word *thupa* refers to a whipping stick. However, on a cow, it is a meat cut that comprises six ribs that are cut in half and separated from the chest ribs (Figure 2, Bone 4 or Figure 1, Part 6). The *thupa* ribs attach to the backbone, which is known as *tlholamatlotla* (Figure 2, Part 3). Once dislodged from *tlholamatlotla*, the *thupa* are added to the *mothobiso* pot which comprises various intestinal meat cuts (cuts from the rumen, small intestines, lungs, heart, omasum) and other small miscellaneous meat cuts. During the wedding celebration, *mothobiso* is eaten at the kgotla by men from both the groom's and the bride's side. It is usually cooked in the kraal or kgotla by those who slaughtered the animal.

No. 7. *Mokoto* (pl. *mekoto*). *Mokoto* is the meat along the spine, from the *mokuana* up to the tail (Figure 1, Part 7). After the *dirope* "hind legs" and the *metlhana* have been removed from both sides of a cow, what is left along the spine is *mokoto*. The butcher separates *mokoto* from *motlhana* at the hip bone (*ischium*). This bone is divided in two so that one part of the bone goes with *motlhana* and the other with *mokoto*. *Mokoto* thus "includes the tail, thoracic vertebrae, lumbar vertebrae, and metapodials" (Mooketsi 2001: 12). *Mokoto* is considered tasty and tender and is cooked in the kraal or in the kgotla. However, during a wedding, it

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1937 (Article)

is eaten at the kgotla, not in the homestead, by all men who attend the celebration. *Mokoto* is considered a male cut and is not eaten by women. According to Ngwato tradition, *mokoto* is cooked with other pieces of meat such as rumen and omasum.

No. 8. *Tshiamo.* In Setswana, the word *tshiamo* means goodness. *Tshiamo* is a piece of meat cut from the tenderloin on the upper side of the cow. It is used in the Ngwaketse traditional wedding ritual of hanging a thin string of omentum around the neck of the bride and groom on their wedding day. In the Bangwaketse doctoring ritual, *tshiamo* is first smeared with traditional medicine by the traditional doctor and then roasted on the hot coals of an open fire. Once cooked, it is handed to the bride's maternal uncle who spikes it on a skewer and feeds it to the couple (Otlogetswe 2020). The couple first bites off small pieces of the doctored *tshiamo*, chews them, and spits them on the ground, after which they can bite, chew, and swallow additional pieces. The meat is believed to have powers to protect the couple against witchcraft (Otlogetswe 2020).

No. 9. *Motlhana* (pl. *metlhana*). This is a piece of meat that is cut after the legs have been removed. It includes the last two ribs of the cow, which have been cut in half, and a substantial portion of meat that stretches towards the tail and includes the hipbone or ischium (Figure 1, Part 9). Two *metlhana* are removed from a *cow*, one from each side, and each has a specific cultural role during a Tswana wedding. *Motlhana* that is removed from one side of the cow is given raw to the bride's maternal uncle, his wife, and the bride's paternal uncle and is then cooked by the bride's maternal aunt. The *sekgawane*, as it is then called, is eaten by the bride and the groom, their parents, and all village elders who are principal players at the wedding. *Sekgawane* is usually served by the bride's maternal uncle to married persons, and the couple will share a plate as a sign that they are one. The second *motlhana* is stored by the bride's family, and carried by her maternal uncle's wife when the bride is delivered to her in-laws following the wedding celebration. The day after the delivery of the bride, tea is prepared and this second *motlhana* is cooked, with rice or samp, and served to the parents and relatives of the groom.

No. 10. *Bobadu*. This cut is found at the rump point before the oxtail. It is also known as *manoka* "the waist meat" since it is found around the centre of the cow. *Bobadu* is eaten together with *mokoto* by the men who have accompanied the groom during the wedding. This meat is eaten by adults only and is rarely given to young people.

No. 11. *Serope* (pl. *dirope*). The word *serope* means a thigh, but in this context refers to the meat of the hind leg. *Serope* is given to the bride's paternal aunt (*rakgadi*), who in traditional Tswana culture is responsible for the sexual education of the bride. Historically, Tswana couples did not always choose their own partners but depended on their parents to make the choice. The betrothed daughter would still be young and inexperienced. To make sure that the marriage was consummated, the bride's paternal aunt would sleep in the same room as the newlyweds on their first night together. If the bride resisted the groom, the paternal aunt was supposed to rebuke her and hold her down. Although this does not occur anymore, the *serope* still carries the nickname of *marage* "the kicker" for the way the new bride would kick the paternal aunt in her struggle with her new husband.

The dictionary should not only account for the beef cuts and their cultural functions. It ought to account for the cattle skeleton, demonstrating a deep cultural appreciation of a cow skeleton as demonstrated in Figure 2.

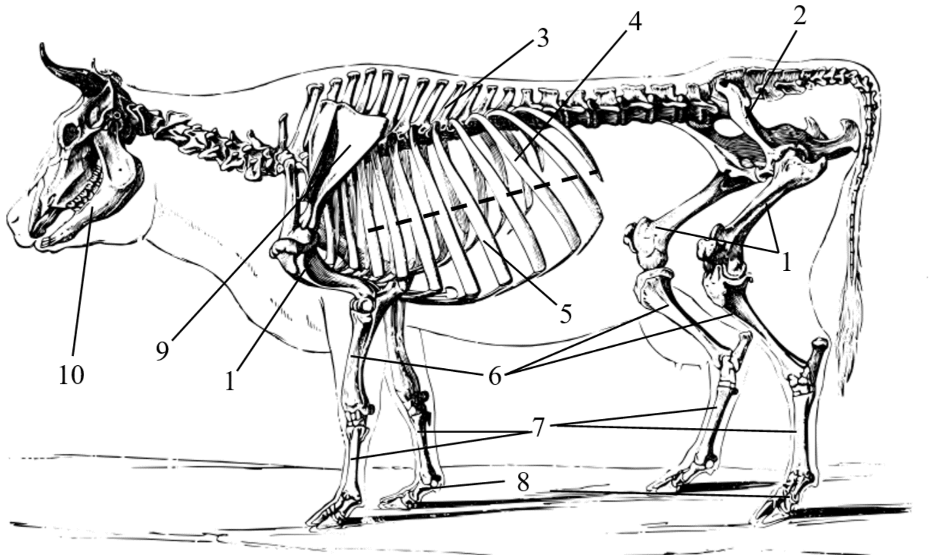


Figure 2: Cow skeleton (Adapted from Otlogetswe 2020)

Such an image could be accompanied by detailed text with the following explanations:

No. 1. *Lesuhu* (pl. *masuhu*). This comprises the femur from the hind leg and the humerus from the front leg of the cow. The humerus links up with the shoulder bone (Figure 2, Bone 9) at the top and the radius and ulna "*mosetlatshe*" at the bottom (Figure 2, Bone 6). The femur links to the hip bone (the ischium) at the top and the tibia and fibula "*mosetlatshe*" at the bottom. *Lesuhu* is cooked whole with small pieces of meat attached to it. The meat pieces are usually eaten after scraping them off the bone with a pocketknife (Mooketsi 2001: 112). The *mašetla*, the proximal ends of the bones, are juicy on the inside and are usually chopped with an axe and chewed. The femur is then cracked open with an axe, either along the shaft or in half, to extract the marrow, which is eaten. During a

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1937 (Article)

Setswana wedding *lesuhu* is eaten by men on the day after the wedding celebration.

No. 2. *Lerapo la noka* "the ischium." It is cut into two with an axe. One half forms part of the *motlhana* (see above) and the other of the *mokoto*.

No. 3. Is *tlholamatlotla* which is the backbone and includes the thoracic, lumbar, and sacral vertebrae. This bone has little meat. It is cut into two and added to the *mothobiso* pot. It is served on the lid to the old men. This bone is believed to be an aphrodisiac.

No. 4. *Thupa*. This comprises the six ribs that are cut in half with an axe to separate them from the chest ribs. Before they are chopped with an axe, the *thupa* ribs attach to the backbone. *Thupa* is added to the pot of *mothobiso* and cooked.

No. 5. *Sehuba*. These are the chest bones that form part of the *sehuba* meat cut that is eaten by married couples.

No. 6. *Mosetlatshe* (pl. *mesetlatshe*). *Mosetlatshe* includes the tibia and fibula in the hind leg and the radius and ulna in the front leg. During a Tswana wedding, these two bones cooked in the same pot as the *masuhu* and are eaten only by women, usually on the day following the wedding celebration in the homestead. Like *masuhu*, these bones are cooked with small pieces of meat attached to them. Once cooked, the pieces of meat are scraped off the bone with a pocketknife and eaten. The bone is then broken with an axe and the extracted bone marrow is also eaten.

No. 7. *Motwane* (pl. *metwane*). *Motwane* is a general term that refers to the metatarsal and metacarpal bones found between the phalanx "*basimanyana ba tlhako*" (Figure 2, Bone 8) and the *mosetlatshe* [tibia, fibula, radius, and ulna] (Figure 2, Bone 6). *Motwane* is eaten on the day that the cow is slaughtered by the men and young men skinning the cow. It is cooked with small miscellaneous meat cuts in the kraal or at the kgotla as part of *mothobiso*. It is of low social status and not highly regarded.

No. 8. *Basimanyana ba tlhako*. These are the proximal and middle phalanx bones (Figure 2, Bone 8) that are usually eaten with the *motwane* [metatarsal or metacarpal bones] or the *tlhako* "hoof" by boys and young men who help with the skinning and butchering of a cow. Weeks, months, and even years after they have been eaten and discarded, they can be found in village playgrounds, pushed around by small children as toys.

No. 9. *Legetla*. This is the scapula [shoulder blade] (Figure 2, Bone 9). During the butchering, it remains attached to the front leg when cut from the carcass. It is therefore part of the meat that pays a traditional doctor or that is cooked by the bride's family during the day *bogadi* is received.

No. 10. *Tlhogo*. This is the skull "*logata*" and the cervical vertebrae of the Tswana (Figure 2, Bone 10). It is eaten by the maternal uncle of the groom and his family. The head together with the neck and the first four ribs from the neck (two on the right and two on the left) are given to the maternal uncle of the groom wrapped in a skin in recognition of his role as head of the wedding negotiations. The maternal uncle takes the *tlhogo* to his house to cook and eat with his family.

A people's culture also includes their food and drink. The depiction of food in the dictionary should be culturally sensitive and not prefer Eurocentric foods to indigenous ones. For instance, a Setswana general dictionary ought to show that Batswana have their own fruits and berries apart from the Eurocentric grapes, pears, peaches, and apples. These include *mmupudu/mompudu* "Common red milkwood (Mimops zeyheri Sonder.)", *mopenoeng/molalakgaka* "Jacket plum (Pappea capensis Ecklon and Zeyher)", *moretlwa*, "Wild berry, (Grewia flava DC)", *Moretologa wa kgomo/Morotologakgomo* "large sourplum (Ximenia caffra Sonder)", *Moretologa wa podi/morotologapodi*, "Small sourplum (Ximenia Americana L.)", Morula "Marula (Scleroccaria birrea (A. Rich.) Hochst.)," *motlhatswa,* Milk plum "Englerophytum magalismontanum," *Ntoroko/motoroko*, "Prickly pear (Opuntia sp)," *mogorwagorwane* "Strychmos Cocculoides," *Mothwane/Mmilo,* "Vangueria infausta," also known as the wild medlar, *motsotsojane* "Kalahari hardy shrub raisin plant (Grewia retinervis Burret)," *mogwana* "Grewia bicolor," *mogwagwa* "Loganiaceae Strychnos madagascariensis," and *morojwa* "Azanza garckeana." These should not be classified simply as *wild*.

An Afrocentric dictionary should reaffirm that before the advent of missionaries and European travellers Africans cultivated their own crops and enjoyed a variety of vegetables, which they continued to enjoy alongside European spinach and cabbage. For instance, the Setswana vegetables include *thepe* (*Amaranthus thunbergii*), *rothwe/rotho*, "Spider whisp (*Cleome gynadra*)," *Thepelešwane (Amaranthus deflexus)*, and *setlepetlepe (Amarunthus Spinosus*). Batswana have cultivated at least twelve crops which have been critical to their survival. These are captured in Appendix 2 (Obopile and Seeletso 2013: 20). No Setswana monolingual or bilingual dictionary illustrates any of Setswana vegetables and crops.

Not only that, an Afrocentric dictionary must also demonstrate that Africans had and still have alcoholic and non-alcoholic beverages such as *mosukujwane/mosukutswane* tea (Lippia Jovania), *kgomodimetsing* (resurrection plant) (Myrothamnus flabellifolius Welw), *longana* (Artemisia afra), *moritelatshwene, Rammola* leaves, *seswagadi* (Jatropha zeyheri Sond), *mosukudu* (Lippia scaberrima) and alcoholic drinks such as *bojalwa jwa Setswana* (traditional sorghum beer), *bojalwa jwa ila*, *khadi* (made from *segwere* "tuber" of *mogakangaga* "Kedrostis hirtella" plant and brown sugar) and *bojalwa jwa morula* (made from fully ripe morula fruits). Such information must enrich the backmatter or the middle section of the dictionary.

A Setswana dictionary must show that for many years Batswana have enjoyed eating various kinds of insects and worms such as *tsie* "locust", some of which have informed many proverbs and idioms such as *kgetsi ya tsie e kgonwa ke go tshwaraganelwa* "it is easier to achieve much by working together." A comprehensive list of worms that Batswana eat is in Appendix 1 (Bultosa et al. 2020: 9).

The discussion above on the beef cuts and their cultural relevance amongst the Batswana as well as a rich Setswana vocabulary on fruits and berries, vegetables, beverages, edible worms, and common crops is a demonstration of what could be included in the middle sections of an Afrocentric dictionary with detailed illustrations. There is much more that an Afrocentric dictionary can do to reflect a people's vibrant culture such as music, dance, clothes, religion, and many other elements which are definitive of a linguistic group. African lexicographers must not lose the fact that a dictionary reinforces a culture.

3. Conclusion

African lexicographers have a profound opportunity to reshape the narrative of African languages through dictionaries that are truly Afrocentric. This article has explored six vital strategies for making African dictionaries resonate with the heartbeat of the continent. These strategies are not just blueprints; they are catalysts for transformation. They represent a commitment to honouring the rich tapestry of languages that have flourished across the continent for centuries. In each meticulously crafted entry, in every nuanced pronunciation guide, and the thoughtful inclusion of cultural context, African lexicographers have the power to breathe life into African dictionaries. They have the power to amplify voices, bridge divides, and foster understanding. They have the power to unlock doors to knowledge, pride, and self-identity. They must craft dictionaries that are not mere repositories of words, but living, breathing testaments to the vibrant mosaic of Africa culture.

References

Dictionaries

Cole, D.T. 1995. *Setswana — Animals and Plants*. Gaborone: The Botswana Society.

Otlogetswe, T.J. 2012. *Tlhalosi ya Medi ya Setswana*. Gaborone: Medi Publishing.

Other references

Atkins, B.T.S., J. Clear and N. Ostler. 1992. Corpus Design Criteria. *Literary and Linguistic Computing* 7(1): 1-16.

Auerbach, R. 1986. First Steps in Setswana Herpetology. *Botswana Notes and Records* 18(1): 71-90.

Barati, A. and P. Noor. 2011. A Study of Methods for Meaning Discrimination of Homonymous and Polysemous Entries: A Case Study of Monolingual English Dictionaries. *Embracing Challenges in Current Trends. Proceedings of the International Conference on Language and Communication (LANCOMM 2011), 19–20 October 2011, Putrajaya, Malaysia:* 26-30. Kuala Lumpur: Kuala Lumpur Infrastructure University College (KLIUC) School of Communication and Language Studies.

Batibo, H. 1996. Loanword Clusters Nativization Rules in Tswana and Swahili: A Comparative Study. *South African Journal of African Languages* 16(2): 33-41.

Batibo, H.M. 1999. A Lexicostatistical Survey of the Setswana Dialects Spoken in Botswana. *South African Journal of African Languages* 19(1): 2-11.

Beier, C. and L. Michael. 2022. Managing Lexicography Data: A Practical, Principled Approach Using FLEx (FieldWorks Language Explorer). Berez-Kroeker, A.L., B. McDonnell, E. Koller and L.B. Collister (Eds.). 2022. *The Open Handbook of Linguistic Data Management*: 301-314. Cambridge, Massachusetts: The MIT Press.

Bultosa, G., M. Molapisi, N. Tselaesele, R. Kobue-Lekalake, G. Desse Haki, S. Makhabu, B. Sekwati-Monang, E. Seifu and P. Nthoiwa. 2020. Plant-based Traditional Foods and Beverages of Ramotswa Village, Botswana. *Journal of Ethnic Foods* 7(1): 1-15.

Crystal, D. and D. Davy. 1969. *Investigating English Style.* London/New York: Routledge.

De Schryver, G.-M. 2010. Revolutionizing Bantu Lexicography — A Zulu Case Study. *Lexikos* 20: 161-201.

Gangla, L.A. 2001. *Pictorial Illustrations in Dictionaries.* Unpublished M.A. thesis. Pretoria: University of Pretoria. Retrieved from:
https://repository.up.ac.za/bitstream/handle/2263/22862/Complete.pdf.

Gouws, R.H. and D.J. Prinsloo. 2005. *Principles and Practice of South African Lexicography*. Stellenbosch: SUN PReSS, AFRICAN SUN MeDIA.

Granger, S. and M. Paquot (Eds.). 2012. *Electronic Lexicography*. Oxford: Oxford University Press.

Gunnink, H. 2020. Language Contact between Khoisan and Bantu Languages: The Case of Setswana. *Southern African Linguistics and Applied Language Studies* 38(1): 27-45.

Hanks, P. 2012. *Corpus Evidence and Electronic Lexicography*. Granger, S. and M. Paquot (Eds.). 2012. *Electronic Lexicography*: 57-82. Oxford: Oxford University Press.

Honey, J. 1997. *Language is Power: The Story of Standard English and Its Enemies.* London/Boston: Faber and Faber.

Matiki, A.J. and G.N. Ramaeba. 2018. Lexical Borrowability in Setswana. *NAWA Journal of Language & Communication* 12(1) 93-110.

McCann, J. 2001. Maize and Grace: History, Corn, and Africa's New Landscapes, 1500–1999. *Comparative Studies in Society and History* 43(2): 246-272.

Miracle, M.P. 1965. The Introduction and Spread of Maize in Africa. *The Journal of African History* 6(1): 39-55.

Moe, R. 2003. Compiling Dictionaries Using Semantic Domains. *Lexikos* 13(1): 215-223.

Mooketsi, C. 2001. Butchery Styles and the Processing of Cattle Carcasses in Botswana. *Pula: Botswana Journal of African Studies* 15(1): 108-124.

Morton, F. and R. Hitchcock. 2014. Tswana Hunting: Continuities and Changes in the Transvaal and Kalahari after 1600. *South African Historical Journal* 66(3): 418-439.

Nkabinde, A.C. 2003. The Implications of Culture for Dictionaries of the African Languages. *Lexikos* 13: 168-182.

Obopile, M. and T.G. Seeletso. 2013. Eat or not Eat: An Analysis of the Status of Entomophagy in Botswana. *Food Security* 5: 817-824.

Otlogetswe, T.J. 2013. *Foreign Words in Setswana Dictionaries.* Available at:
https://otlogetswe.wordpress.com/2013/08/04/foreign-words-in-setswana-dictionaries/

Otlogetswe, T.J. 2016. Are these Words, Setswana Words? *Sunday Standard*, 3 November. Available at: https://www.sundaystandard.info/are-these-words-setswana-words-2/.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1937 (Article)

Otlogetswe, T.J. 2019. *The Great Marage: How at Night Rakgadi Held Her Legs as the Man Mounted Her*, 7 April 2019. Available at:
https://otlogetswe.wordpress.com/2019/04/07/the-great-marage-how-at-night-rakgadi-held-her-legs-as-the-man-mounted-her/

Otlogetswe, T.J. 2020. Beef Cuts amongst the Bangwaketse: The Case of Motlhakanelwa. *Anthropology Southern Africa* 43(4): 233-245.

Pawliková-Vilhanová, V. 2009. White Fathers' Linguistic Work and Contribution to the Development of African Languages and Literatures. Prah, K. 2009. *The Role of Missionaries in the Development of African Languages*: 61-90. Cape Town: Centre for Advanced Studies of African Society.

Peterson, R.A. 1979. Revitalizing the Culture Concept. *Annual Review of Sociology* 5(1): 137-166.

Prah, K.K. (Ed.). 2009. *The Role of Missionaries in the Development of African Languages*. Cape Town: Centre for Advanced Studies of African Society.

Prinsloo, D.J. 2009. Current Lexicography Practice in Bantu with Specific Reference to the *Oxford Northern Sotho School Dictionary*. *International Journal of Lexicography* 22(2): 151-178.

Prinsloo, D.J. 2015. Corpus-based Lexicography for Lesser-resourced Languages — Maximizing the Limited Corpus. *Lexikos* 25: 285-300.

Prinsloo, D.J. 2017a. Analyzing Words as a Social Enterprise: Lexicography in Africa with Specific Reference to South Africa. Miller, J. (Ed.). 2017. *Analysing Words as a Social Enterprise: Celebrating 40 Years of the 1975 Helsinki Declaration on Lexicography*. Available online at:
https://www.adelaide.edu.au/australex/publications/

Prinsloo, D.J. 2017b. 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.

Prinsloo, D.J. and N. Zondi. 2020. From Postcolonial African Language Lexicography to Globally Competitive e-Lexicography in Africa. Kaschula, R.H. and H.E. Wolff (Eds.). 2020. *The Transformative Power of Language: From Postcolonial to Knowledge Societies in Africa*: 259-274. Cambridge: Cambridge University Press.

Reynierse, C. (Ed.). 1996. *South African Multi-language Dictionary and Phrase Book.* Cape Town: Reader's Digest Association of South Africa.

Robbins, K. (Ed.). 2017. *The History of Oxford University Press: Volume IV: 1970 to 2004.* Oxford: Oxford University Press.

Sands, B. and K. Jones. 2022. *N|uuki Namagowab Afrikaans English Dictionary*. Stellenbosch: African Sun Media.

Southerland, R.H. and F. Katamba. 1997. Language in Social Contexts. O'Grady, W., M. Dobrovolsky and F. Katamba (Eds.). 1997. *Contemporary Linguistics: An Introduction:* 540-590. London: Longman.

Spencer-Oatey, H. 2012. *What is Culture? A Compilation of Quotations. GlobalPAD Core Concepts.* Available at GlobalPAD Open House: http://go.warwick.ac.uk/globalpadintercultural

Svensén, B. 2009. *A Handbook of Lexicography. The Theory and Practice of Dictionary-making*. Cambridge: Cambridge University Press.

Trench, R.C. 1857. *On Some Deficiencies in Our English Dictionaries: Being the Substance of Two Papers Read before the Philological Society, Nov. 5, and Nov. 19, 1857.* London: John W. Parker and Son.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1937 (Article)

Tylor, E.B. 1871. *Primitive Culture: Researches into the Development of Mythology, Philosophy, Religion, Art, and Custom. Vol. 2.* London: John Murray.

Van Wyk, E.B. 1995. Linguistic Assumptions and Lexicographical Traditions in the African Languages. *Lexikos* 5: 82-96.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1937 (Article)

Appendix 1: Tswana edible worms

Setswana name	**English name**	**Scientific name**
Phane	Mopane worm	*Imbrasia belina* Westwood
Sega/Sengana	Arrow sphinx	*Lophostethus dumolinii* Angas
Khana	Oleander hawk moth	*Daphnis nerii* L.
Phata	Common emperor moth	*Bunaea alcinoe* Stoll
Kokobele	Harvester termite	*Hodotermes mossambicus* (Hagen)
Sekala	Marbled emperor moth	*Heniocha* spp.
Tutukgweba/Lebitse	Giant jewel beetle	*Stenorcera orissa* Buq.
Ntlhwa	African thief ant	*Carebara vidua* F. Smith
Tsie molome	Red locust	*Normadacris septemfasciata* Serville
Tsie segongwane	Brown locust	*Locustana pardalina* Walker
Tsie ya sekaka	Desert locust	*Schistocerca gregaria* Forskal
Nato	Pallid emperor moth	*Cirina forda* Westwood
Kgonono	Willow emperor moth	*Imbrasia tyrrhea* Cramer
Dikakabotha	Sundown emperor moth	*Sphingomorpha chlorea* Cramer
Thethe	Scarab larvae	*Oryctes boas* Fabr.
Tsiakgope	Brown-spotted locust	*Cyrtacanthacris tatarica* L.
Lentloro	Common stick grasshopper	*Acrida acuminata* Dirsh.
Morwerwe	Silver striped hawk	*Hippotion celerio* L.
Notshe	Honeybee	*Apis mellifera* L.
Senyetse	Cicada	*Monomatapa insingnis* Distant
Ntsi ya mooka	Stingless bee	*Plebeina hildebrandti Friese*
Kokomochane	Stingless bee	*Hypotrigona gribodoi* Magretti,
Tsie ya matebele	Elegant grasshopper	*Zonocerus elegans* Thunb.
Mmamati	Burrowing grasshopper	*Acrotylus spp.*
Mookotsane	Mopane bees	*Meliponula sp*
Senyanyantsodi	Speckled emperor moth	*Gynanisa maja* (Klug)
Monakamongwe	Convolvulus hawk moth.	*Agrius convolvuli* L.

Adapted from Bultosa et al. (2020: 9).

Appendix 2: Setswana common crops

Setswana	**English**	**Scientific Name**	**Parts Used**
Mabele	Sorghum	*Sorghum bicolor L.*	Grains
Mmidi	Maize	*Zea mays L.*	Grains
Lebelebele	Pearl millet	*Pennisetum glaucum (L.) R. Br.*	Grains
Dinawa	Common beans	*Phaseolus vulgaris L.*	Pulses/seeds
Ditloo	Jugo beans	*Vigna subterranean (L.) Verdc*	Pulse/seeds
Letlhodi	Mung bean	*Vigna radiata (L) R. Wilczek*	Pulse/seeds
Manoko/matonkomane	Groundnuts	*Arachis hypogaea L.*	Nuts
Lephutshe	Pumpkin	*Cucurbita pepo L.*	Pulp
Magapu	Watermelon	*Citrullus lanatus (Thunb.) var. lanatus*	Pulp & seeds
Makgomane	Squash	*Lagenaria siceraria (Molina) Standl.*	Fruit
Marotse/makatane/ maowane	Melon	*Citrullus lanatus var. citroides (L.H. Bailey) Mansf.*	Pulp, seeds, rind
Ntšhê	Sweet reed	*Sorghum bicolor (L.) Moench*	Stem reed

Adapted from Obopile and Seeletso (2013: 20).

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1940

A Scoping Review of Studies into Dictionary Use and Language Learning

Xiaoshuai Ge, *School of Foreign Languages,*
Shandong Agricultural University, Tai'an, China; and
Center for Linguistics and Applied Linguistics,
Guangdong University of Foreign Studies,
Guangzhou, China (gexiaoshuai89@foxmail.com)
(https://orcid.org/0000-0003-4978-3934)

Songshan Zhang, *Center for Linguistics and Applied Linguistics, Guangdong*
University of Foreign Studies, Guangzhou, China
(Corresponding author, alex_zhang@gdufs.edu.cn)
(https://orcid.org/0000-0003-2033-4561)

Hai Xu, *Center for Linguistics and Applied Linguistics, Guangdong*
University of Foreign Studies, Guangzhou, China
(xuhai1101@gdufs.edu.cn) (https://orcid.org/0000-0003-4644-9033)
and
Xian Zhang, *Department of Linguistics, University of North Texas,*
Denton, USA (xian.zhang@unt.edu)
(https://orcid.org/0000-0001-8472-5380)

Abstract: This study presents a scoping review of empirical studies on dictionary use and language learning, with the aim to systematically examine the development and trends of the field across a specific timespan. Based on the content analysis of 104 journal articles published between 1992 and 2024, this study has several major findings. The analysis of publication trends indicates that there is a surge in research interest after 2010, largely due to the rise of digital technology and the increasing recognition of the role of dictionaries in self-directed learning. While the "Big Six" dictionaries are widely used across studies, their electronic counterparts have evolved as the dominant dictionary form, mirroring the impact of the Digital Revolution. It is revealed that there is a growing presence of mixed-method design studies, reflecting the field's increasing concern for both the learning outcome induced by dictionary use and learners' cognitive processes involved in dictionary look-up behavior. It is also found that there is an overreliance on advanced and intermediate EFL/ESL learners to generate conclusions, underscoring the need for studies involving learners of diverse target L2s and varied L2 proficiency levels. Surprisingly, despite the huge potential evinced by electronic dictionaries to support language learning in areas like grammar, pragmatics, and pronunciation, research in these domains remains limited. This scoping review underscores the need for further research, especially those conducted with longitudinal design and in naturalistic contexts, to foster a holistic understanding of how dictionaries can enhance language learning.

Lexikos 34 (AFRILEX-reeks/series 34: 2024): 331-354

Keywords: DICTIONARY, DICTIONARY USE, DICTIONARY LOOKUPS, LANGUAGE LEARNING, LANGUAGE ACQUISITION, SCOPING REVIEW, SECOND LANGUAGE, EFL/ESL LEARNERS

Opsomming: 'n Evaluerende oorsig van studies oor woordeboekgebruik en taalleer. In hierdie artikel word 'n evaluerende oorsig van empiriese studies oor woordeboekgebruik en taalleer aangebied, met die doel om die ontwikkeling van die veld asook die tendense in die veld oor 'n spesifieke tydperk, sistematies te ondersoek. Gebaseer op die analise van die inhoud van 104 joernaalartikels wat tussen 1992 en 2024 gepubliseer is, word daar verskeie belangrike bevindings in hierdie studie gemaak. Die ontleding van publikasietendense dui daarop dat daar 'n oplewing in navorsingsbelangstelling ná 2010 is, grootliks as gevolg van die opkoms van die digitale tegnologie en die toenemende erkenning van die rol van woordeboeke in selfgerigte leer. Terwyl die "Groot Ses"-woordeboeke wyd in studies gebruik word, het hul elektroniese eweknieë ontwikkel in die dominante woordeboekvorm, wat die impak van die Digitale Revolusie weerspieël. Daar word aan die lig gebring dat ontwerpstudies wat gemengde metodes behels, al meer voorkom, wat toenemende besorgdheid in die veld weerspieël oor sowel die leeruitkoms wat deur woordeboekgebruik teweeggebring word as die leerders se kognitiewe prosesse wat by woordeboeknaslaangedrag betrokke is. Daar is ook bevind dat daar oormatig gesteun word op gevorderde en intermediêre EVT-/ETT-leerders om resultate te verkry, wat die behoefte aan navorsing wat leerders van uiteenlopende tweedetaaldoeltale en gevarieerde tweedetaalvaardigheidsvlakke betrek, beklemtoon. Ten spyte van die groot potensiaal van elektroniese woordeboeke om taalleer in domeine soos die grammatika, pragmatiek en uitspraak te ondersteun, bly navorsing op hierdie gebiede, verrassend genoeg, beperk. Hierdie evaluerende oorsig beklemtoon die behoefte aan verdere navorsing, veral dié wat met longitudinale ontwerp en in natuurlike kontekste uitgevoer word, om sodoende 'n holistiese opvatting te bevorder van hoe woordeboeke taalleer kan verbeter.

Sleutelwoorde: WOORDEBOEK, WOORDEBOEKGEBRUIK, WOORDEBOEKNASLAANPOGINGS, TAALLEER, TAALVERWERWING, EVALUERENDE OORSIG, TWEEDE TAAL, EVT-/ETT-LEERDERS

1. Introduction

Nearly half a century ago, Wiegand had called on lexicographers to pay attention to knowledge about users gained from empirical studies "to write more adequate dictionary entries" (cf. Welker 2013: 532). However, "the take-off was very slow" (Tarp 2009). As noted by Nesi (2014), the increased interest in dictionary use research (see Atkins 1998) was to a very large extent fueled by the advent of 'learners' dictionaries' in the 1980s and 1990s, which has reshaped the dictionary content and design. Fast forward to the 21st century, the use of dictionaries has evolved. Benefited from the Digital Revolution (Fellbaum 2014; L'Homme and Cormier 2014; Lew and De Schryver 2014), users now employ a variety of dictionary forms — print, electronic, online, and dictionary applications — to assist their learning of both native and foreign languages. This also has triggered a new wave of dictionary use research, with a focus on electronic dictionaries (Lew and De Schryver 2014; Müller-Spitzer 2014).

Dictionary use research covers many sub-fields, and researchers seem to have not reached a consensus on its scope. For example, Hulstijn and Atkins (1998) outlined seven topics of dictionary use research while Nesi (2014) identified five recurring issues of dictionary use research. Different as their categorizations are, they all seem to agree that the relationship between dictionary use and L2 learning is a central issue in dictionary use research, as also observed by Hartmann (2001). As pointed out by Hulstijn and Atkins (1998), dictionary use itself is a complex and subtle activity, and investigating the effects of dictionary use on L2 learning involves many variables, including dictionary user-related variables (users' sophistication, proficiency, capability of understanding the metalanguage, familiarity with the target dictionary, and knowledge of the subject matter), task-related variables (task format and type), variables relating to both dictionary users and learning tasks (task difficulty, target L1 and L2, and type of linguistic unit), and dictionary-related variables (dictionary form, dictionary type, way of information presentation, source of information available, and adequacy of coverage). However, studies into dictionary use and language learning seem to be fragmentary and no study has ever attempted to systematically synthesize the findings, although there exist several preliminary brief reviews that touch upon certain aspects of this topic.

Literature reviews play a crucial role in guiding and benefiting researchers within a given field by summarizing and synthesizing existing knowledge. They come in various forms, including traditional literature reviews, scoping reviews, meta-analyses, and annotated bibliographies, each serving a distinct purpose in academic discourse. The present study aims to explore the interplay between dictionary use and language learning by employing a scoping review methodology.

A scoping review is a systematic and iterative approach to knowledge synthesis, designed to map the existing or emerging literature on a specific topic (Mak and Thomas 2022). It is particularly adept at providing researchers with a rapid overview of the main areas of interest and identifying significant gaps in the literature (Arksey and O'Malley 2005). Since its inception into the academia, numerous scoping reviews have been conducted in fields such as healthcare and education. This study attempts to adopt the scoping review methodology to offer a systematic and comprehensive review on the subject of dictionary use and its impact on language learning. Specifically, our investigation will be guided by the following questions:

— What research methods and research instruments were used to investigate the relationship between dictionary use and language learning in the selected studies?

— What trends could be revealed by the study distribution across publication year, academic journals, and study location?

— What languages and language learners were covered in the selected studies? Are learners with diverse L2s and varied L2 proficiency levels well-represented in the sampled studies?

— What dictionaries were sampled as the subject of investigation in the selected studies? What forms and types of dictionaries have been the interest for research over time?

— What are the research foci of the selected studies?

2. Literature review

As mentioned in the introductory section, there have been several reviews conducted by scholars in the field (e.g., Hulstijn and Atkins 1998; Lew 2011; McCreary and Doležal 1998; Nesi 2014; Töpel 2014), but unfortunately, they failed to provide answers to the above-mentioned questions due either to the depth of analysis, analytical procedures, or the scope of investigation. In what follows, we will provide a brief review of these studies.

In the domain of lexicography, a notable tradition exists for creating annotated bibliographies. Hulstijn and Atkins (1998) contributed to this tradition by compiling an annotated bibliography encompassing approximately 50 publications focused on dictionary use and foreign language learning. Following in these footsteps, Doležal and McCreary (1999) assembled an annotated bibliography consisting of 521 publications on pedagogical lexicography, with a particular emphasis on language learners and dictionary users. Additionally, Welker's *O Uso de Dicionários: Panorama Geral Das Pesquisas Empíricas* provides a comprehensive annotated bibliography of 220 publications on empirical studies related to dictionary use (cf. Lew 2007). These works cover a wide range of studies involving multiple European languages, such as English, German, and Portuguese. Also, they often feature chapters that categorize the literature, highlight research gaps, and offer directions for future research. In a similar vein, the article by Nesi (2014) and the book chapter by Töpel (2014) could also be regarded as annotated bibliographies, with Nesi listing 35 publications on dictionary use by English learners and Töpel (2014) briefly reviewing 35 studies on the use of electronic dictionaries. Admittedly, these above-mentioned studies have touched upon various aspects of dictionary use research. However, they all did not concentrate the focus on the relationship between dictionary use and language learning. For example, among the five topics on dictionary use identified by Nesi (2014), only one was concerned with dictionary use and language learning, indicating that only a small number of studies was covered on this topic and that the scope of investigation was rather limited. Similarly, Töpel's (2014) review only covered several studies on electronic dictionary use and language learning. Also, more often than not, these surveys (e.g., Hulstijn and Atkins 1998; Doležal and McCreary 1999; Nesi 2014) only presented a brief summary of the

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1940

included studies, and failed to offer an in-depth and systematic content analysis. One more limitation of theses studies is that they were not conducted by following standard procedures of scoping reviews or systematic reviews as they usually did not specify the literature retrieval procedures and coding schemes, making them not replicable.

Likewise, in the special issue "Studies in Dictionary Use: Recent Developments" of the *International Journal of Lexicography*, Lew (2011) authored an introduction to the studies within the issue and offered critical evaluations of the trends of dictionary use research. Illuminating as Lew's (2011) work was, it was also confined to only six studies on dictionary use included in the special issue. Interestingly, a review by McCreary and Doležal (1998) exhibited similarities to the more recent scoping reviews. In their proceedings paper, they reported on the trends and findings from a previous bibliography of 460 publications, signaling an early adoption of systematic review methodologies and the identification of research gaps within a focused topic, despite not adhering to modern scoping review protocols. However, few empirical studies into dictionary use and language learning were included, probably due to the fact that at the time of conducting the review, studies on this topic were rather limited in number.

In addition, there were two narrative reviews on vocabulary acquisition through dictionary use under intentional/incidental learning conditions (Ronald 2003a, 2003b) and a list of studies on dictionary-induced vocabulary learning (Welker 2010). However, like the most recent meta-analytic review in the field of dictionary use and vocabulary learning conducted by Zhang et al. (2021), which synthesized findings from 44 studies and examined the effects of dictionary use on second language vocabulary acquisition, the scope of these reviews was limited to vocabulary only, failing to cover other aspects of language learning.

To sum up, the review above reveals that there is a lack of systematic research synthesis and content analysis on the relationship between dictionary use and language learning. Given that scoping review has become a matured and useful method in helping researchers map out the landscapes of a research domain and that numerous fruitful outcomes have been yielded by using this methodology in areas such as health care and education, we attempt to present a scoping review of studies into dictionary use and language learning. In what follows, we will be guided by the research questions listed at the end of the Introduction section and present the procedures for conducting the scoping review, the generated results, the implications of the findings, and the conclusions drawn from the review.

3. Method

According to the framework proposed by Arksey and O'Malley (2005), there are five procedures to follow in conducting a scoping review, namely formulating the research questions, locating potential studies, selecting target studies, charting the data, and summarizing and reporting the data. Therefore, these steps were strictly executed to ensure that this study was methodologically transparent and replicable, and that the findings generated were reliable. As we have listed the research questions at the end of Section 1, we will outline the next steps regarding research methodology here.

3.1 Literature retrieval procedures

We firstly consulted Web of Science, Education Recourses Information Center (ERIC), and ProQuest Linguistics and Language Behavior Abstracts (LLBA) to locate relevant studies. These databases were chosen because they had a broad coverage and were frequently used in research synthesis. As the focus of this study was on dictionary use and language learning, two sets of keywords were used to launch the database search: dictionary-related keywords, including dictionary, monolingual dictionary, bilingual(ized) dictionary, electronic dictionary, paper dictionary, online dictionary, dictionary use, as well as language learning-related keywords, including vocabulary, grammar/syntax, pragmatic knowledge, reading, writing, and translation. Then, key academic journals in the field of lexicography, such as *Dictionaries*, *Lexicographica*, *Lexicography*, were manually searched to retrieve studies that were not identified through electronic database search. In addition, we also browsed the references of relevant books, book chapters, reviews, and research articles to trace potential studies.

3.2 Inclusion criteria

Five inclusion criteria were specified as the filters for appropriate articles selection. The following requirements had to be met if an article was to be considered for inclusion in this scoping review:

- It should be published before February 6th, 2024, which was the cutoff date for data collection;
- It should be reported in the English language, given that English is a de facto international lingua franca and is also prevalent in international lexicographical discourse;
- It should investigate language learning through dictionary use by language learners;
- It should provide empirical evidence about the effectiveness of dictionary use on learners' learning outcomes.
- It should explicitly report the information needed for data coding (See Section 3.3).

Articles were excluded if it was not published within the cutoff date or not reported in English. Publication types such as literature reviews, book reviews, or editorials were also excluded for inclusion. Empirical studies that only investigate dictionary users' needs, learners' dictionary lookup behaviors, strategies or skills were not included, as they do not include language learning achievements as the outcome variable and the focus of this review was on empirical studies investigating the effects of dictionary use on language learning.

3.3 Coding scheme

Following the practices of previous scoping reviews (Hung et al. 2018; O'Flaherty and Phillips 2015) and in reference to our research questions, we classified the variables into five categories: publication-related variables, treatment-related variables, methodology-related variables, outcome-related variables, and learner-related variables.

3.3.1 Publication-related variables

This category contains four variables describing the meta-information of each selected article: study ID (the identification number of the study), author (the author of the study), year of publication (the year in which the study was published), and publication type (journal article, book chapter, or dissertation).

3.3.2 Treatment-related variables

This category mainly concerns dictionaries used, dictionary form, dictionary type, and research setting (laboratory or classroom). We firstly identified the dictionaries used in each study. As it would be cumbersome to list all individual dictionaries, dictionaries belonging to the same family or series were listed under one category, such as the Oxford series. Dictionary form was categorized as electronic or paper, while dictionary type consisted of three subcategories: monolingual, bilingual, or bilingualized. For research setting, studies were coded as laboratory or classroom depending on the location where they were conducted.

3.3.3 Methodology-related variables

This category included research methods, research design, research instruments, and type of assessment. Research methods were divided into the quantitative method (involving statistical analyses), the qualitative method (no statistical analyses), and the mixed method (involving both quantitative data and qualitative data) by following the practices of similar scoping reviews (e.g., Hung et al. 2018),

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1940

while research design referred to whether a study adopted a between-groups design, a within-group design, or a correlational design, or whether it was observational or self-retrospective in nature. Research instruments denoted the specific measurement tools used in the study, such as vocabulary tests and eye-tracking technique. Assessment type was coded as productive test (usually requiring learners to produce language structures or texts), receptive test (usually in the form of multiple-choice tests or yes/no tests), or mixed.

3.3.4 Outcome-related variables

Outcome-related factors consisted of learning outcome and result orientation. The former referred to learners' learning outcomes. Following Stockwell (2007), it was coded as the macro skills of reading, writing and translation, as well as vocabulary, grammar, pronunciation. Results orientation indicated whether the study yielded a positive effect of dictionary use or not. Therefore, it was coded as positive, negative, no significance, or mixed.

3.3.5 Learner-related variables

This group of variables encompassed learners' educational level, proficiency, educational context, L1, and L2. Learners' educational level was coded as kindergarten, primary school, secondary school, university, or mixed. Proficiency was defined as low, intermediate, advanced, or mixed. Educational context was specified as foreign language context or second language context depending on whether learners learned a foreign language in a country or region where the target language is a foreign language, or the official language/second language. Learners' L1 and L2 were coded as their first language or second language, respectively.

3.4 Data collection and coding reliability

The procedures for literature retrieval, screening, and selection were presented in Figure 1 below. For all the 2707 articles identified through automatic database search, a researcher first read through the abstracts to make a judgment of their eligibility. This initial screening procedure culled out 2472 ineligible articles as they either failed to meet the inclusion criteria or did not directly address the research questions of interest, leaving 235 potentially useful ones. Then two researchers worked collaboratively through careful reading of the full texts to finalize the list of publications eligible for inclusion. In addition, three articles gleaned from the reference lists of relevant publications but not identified through database search were also included. Table 1 shows the distribution of articles across academic journals.

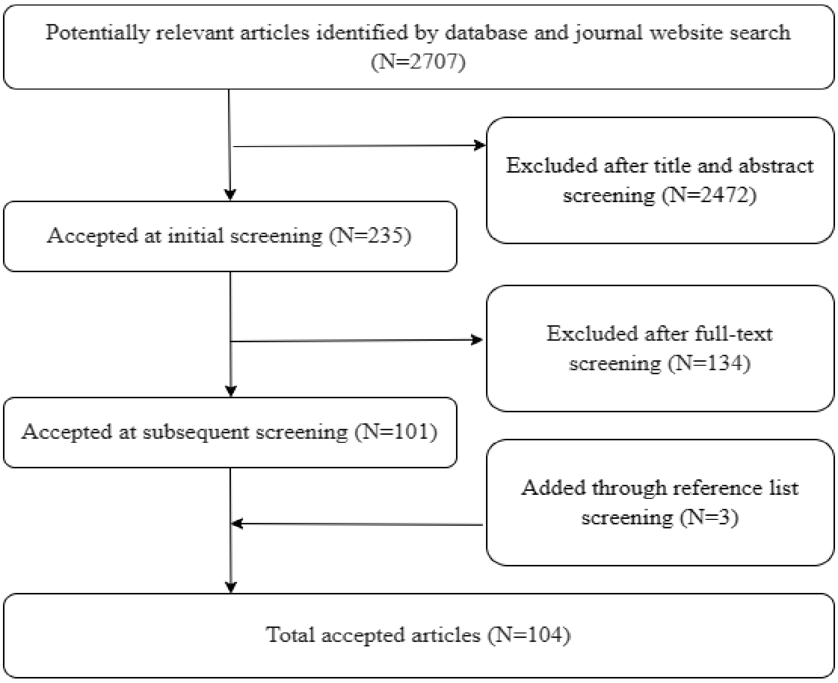

Potentially relevant articles identified by database and journal website search (N=2707)
Excluded after title and abstract screening (N=2472)
Accepted at initial screening (N=235)
Excluded after full-text screening (N=134)
Accepted at subsequent screening (N=101)
Added through reference list screening (N=3)
Total accepted articles (N=104)

Figure 1: Flow chart for data acquisition

Two researchers were involved in coding the articles. At first, five articles were randomly selected and double-coded by the two coders. This was done to make sure that a tentative agreement on how each factor should be coded could be reached. Thereafter, one researcher undertook to code all the remaining articles and checked the results multiple times to maximize intra-coder reliability. Upon completion, twenty-five articles, accounting for almost one fourth of the remaining total, were selected and handed over to a second researcher for coding. The agreement ratio for all values in the coding sheet among the two coders was 91%. The observed discrepancies between the two coders were resolved through discussion. For example, as there is a lack of unified criteria for judging learners' proficiency level, one coder would rely on his own judgement and code learners' proficiency level as intermediate if they were from a middle school, while the other coder would code it as advanced if they were from a higher proficiency group or low if they were from a lower proficiency group based on the researcher's report, even if they were all middle school learners. The two coders then discussed about this issue and reached a consensus that learners' proficiency level should be coded based on the researcher' report, following the practice of similar scoping reviews, systematic reviews, or meta-analytic reviews. Finally, the overall results were re-examined thoroughly by the first coder to eliminate potential inconsistencies.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1940

3.5 Data analysis

According to Arksey and O'Malley (2005), the next stage is to analyze the data and a content analysis approach (cf. Krippendorff 2018) was adopted to do so. Summaries for each article were generated in terms of the 16 variables, subsumed under the five categories as outlined in Section 3.3. All the data were imported into Microsoft EXCEL and SPSS 22.0, and we mainly adopted frequency counts to describe the data and *t*-test for inferential statistical analysis. Data summaries and results were presented in the section that follows.

4. Results and discussion

4.1 Research methodology and research instruments employed in selected studies

In terms of research methods, it was revealed that all the sampled studies involved some sorts of statistical analysis, therefore they were all quantitative in nature. As for research design, the analysis of the 104 selected studies showed that a preponderance (n=90, 87%) was found to utilize experimental approaches, nearly half (n=44, 42%) incorporated survey techniques, and a significant minority (n=26, 25%) engaged observational methodologies. It is noteworthy that the aggregate number of methodological instances surpasses the total count of studies, a discrepancy attributable to the fact that 43 studies incorporated mixed-method designs.

We use 'experimental' as an umbrella term for both experimental and quasi-experimental designs. It is observed that numerous studies within the sample refer to 'experiments' in contexts that are more accurately described as quasi-experiments, particularly when the allocation of subjects to treatment and control groups lacks randomization. The research instruments commonly employed in these experimental studies include vocabulary retention tests, recall tests, fill-in-the-blank assessments, and matching tests.

Survey methods include questionnaire studies and interviews. Questionnaires and interviews are the instruments frequently adopted before or after the experiment to collect the subjects' attitudes toward the experiment conditions. They are also employed independently in survey studies. In contrast, observational studies are less frequently adopted in dictionary use studies. The instruments used for observation are screen-recordings, eye-trackers, search records, logfiles, and think-aloud recordings, each offering unique insights into the behaviors and processes of dictionary use.

Within the field of dictionary use, it is somewhat unexpected to find a scarcity of observational studies. While sophisticated tools like eye-trackers necessitate both laboratory settings and specialized expertise for effective use, alternative methods such as screen recordings and think-aloud protocols are more

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1940

accessible yet remain underutilized. These latter instruments, despite their relative simplicity, have the potential to yield rich and valuable data when applied thoughtfully in research. For example, Chen and Liu (2022) leveraged the utility of screen recording to gain insights into the use of the Bing bilingual dictionary within the context of EFL writing. This approach facilitated a granular examination of the interactions between users and the dictionary interface. Similarly, Kim (2018) employed the think-aloud protocol to evaluate the effectiveness of teaching English article rules and to understand how dictionary consultation could improve the usage of articles among learners.

Another critical aspect to consider is the temporal scope of the studies. An overwhelming majority, constituting 87% of the total (n=90), are characterized as cross-sectional in design. In contrast, a considerably smaller subset, representing only 13% (n=14), employs a longitudinal approach. This discrepancy underscores an area that merits greater attention from researchers in future endeavors. The relative scarcity of longitudinal studies suggests an opportunity for longitudinal analyses that could offer more comprehensive insights into the long-term effects of dictionary use on language learning.

4.2 Distribution of studies across publication year, academic journals, and study location

As mentioned earlier, there is a total of 104 studies, published between 1992 and 2024, included in this scoping review. Figure 2 presents the publication frequency, namely the number of publications, of studies into dictionary use and language learning in each year. As can be seen from Figure 2, the number of publications before 2010 was relatively low and it remained stable across this time range. However, there was a remarkable increase in the number of publications since 2010, and there had already been two articles published in the first two months of 2024. Descriptive statistics showed that the average annual number of publications before and after 2010 was 1.44 (SD = 1.15) and 5.13 (SD = 2.39), respectively. Results from *t*-test revealed that the difference in terms of publication frequency between the two periods was significant ($F = 6.66$, $t = 5.81$, $p = .02 < .05$). These findings obviously suggest that the area of studies concerning the effectiveness of dictionary use on aspects of language learning has been increasingly gaining scholarly attention over the last decade. By analyzing the publications across the timespan, we can see that the increase in the number of publications after 2010 to a very large extent can be attributed to the rapid development of digital technology and the increasing recognition of the role of dictionaries in self-directed learning. The expansion of research into various aspects of dictionary use reflects a maturing field that is beginning to address more nuanced questions at the intersection of lexicography, language pedagogy, and technology.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1940

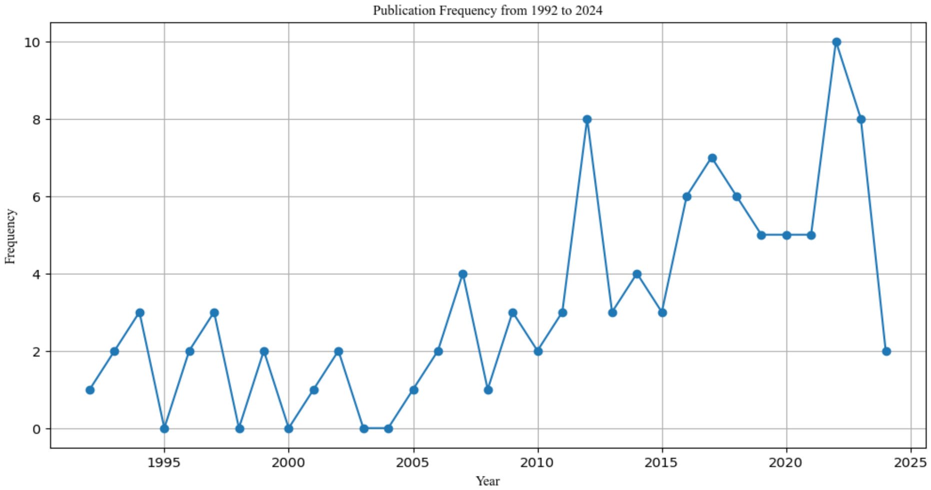


Figure 2: Distribution of studies across year of publication

Presented in Table 1 was the distribution of articles across different academic journals. It can be seen that academic journals that published relatively larger number of studies concerning the effects of dictionary use on language learning over the last three decades were *International Journal of Lexicography*, *Lexikos*, *Computer Assisted Language Learning*, *ReCALL*, and *Computers & Education*, respectively. Among these five journals, *International Journal of Lexicography* published a total of 43 articles, which was the largest number and accounted for 41.3% of all articles included in this scoping review, followed by *Lexikos*, which yielded 14 articles and accounted for 13.5% of all publications. Taken together, articles published in *International Journal of Lexicography* and *Lexikos* took up more than a half (54.8%) of all the included publications. This is perhaps not difficult to understand, as these two journals mainly publish articles in lexicography and its related disciplines, and are the only ones that are indexed by the Arts & Humanities Citation Index and the Social Sciences Citation Index in the field of lexicography. For example, on the homepage of *International Journal of Lexicography*, it is clearly stated that '*… it is concerned with all aspects of lexicography, including issues of design, compilation and use, and with dictionaries of all languages …*' [emphasis added]. Therefore, it can be seen that dictionary use is a key topic in the articles considered for publication in the journal.

However, it is somewhat surprising to see that the other three journals devoted to publishing articles in lexicography, namely *Dictionaries*, *Lexicography*, and *Lexicographica*, only published three articles on dictionary use and language learning, with two in *Lexicography* and one in *Dictionaries*, respectively. This might indicate that theoretical articles or meta-lexicographical/ontological studies dominate these three journals, or it could be because these three journals received relatively lower number of submissions on this topic as they do not boast the far-researching influence and impact when compared with *International Journal of Lexicography* and *Lexikos*.

Table 1: Article distribution across academic journals

Publication Title	Count	Publication Title	Count
International Journal of Lexicography	43	*PLOS ONE*	1
Lexikos	14	*Educational Technology & Society*	1
Computer Assisted Language Learning	7	*Early Child Development and Care*	1
ReCALL	6	*System*	1
Computers & Education	5	*Education and Information Technologies*	1
The Modern Language Journal	3	*Language Testing*	1
Applied Linguistics	3	*Journal of Research on Technology in Education*	1
Lexicography	2	*Journal of Educational Research*	1
Language Learning & Technology	2	*Educational Review*	1
Language Learning	2	*Journal of Adolescent & Adult Literacy*	1
Language Teaching Research	2	*Dictionarie*s	1
Innovation in Language Learning and Teaching	1	*Educational Technology Research and Development*	1
TESOL Quarterly	1	*Applied Psycholinguistics*	1

It is also worth mentioning that studies into dictionary use and language learning also appeared relatively frequently in language learning technology related journals such as *Computer Assisted Language Learning*, *ReCALL*, and *Computers & Education*, or less frequently in *Language Learning & Technology*, *Educational Technology & Society*, *Education and Information Technologies*, *Journal of Research on Technology in Education*, and *Educational Technology Research and Development*. This obviously reflects the impact of the Digital Revolution on lexicography and dictionary user research (Fellbaum 2014; L'Homme and Cormier 2014; Lew and De Schryver 2014), where researchers began to focus on how the use of electronic or online dictionaries might influence language learning outcomes (e.g., Chen 2022; Chen and Liu 2023; Dziemianko 2022; Li and Xu 2015; Lo 2024; Rees and Lew 2024; Tsai 2019). It also mirrors the interdisciplinary nature of lexicography and the pervasive influence of computer science and educational technology on dictionary making and lexicographical studies. This may explain why some early scholars (e.g., Sinclair 1984) even challenged the status of lexicography as an academic subject and placed it at the crossroads of linguistics and information technology. The importance of computer science and information technology to lexicography is also reflected in the aims and scopes of academic journals in the field. For example, the publication policy of *Lexikos* states that '*articles dealing with all aspects of lexicography or the implications that research in related disciplines such as linguistics, computer and information science, etc. has for lexicography will be considered for publication*' [emphasis added].

Figure 3 presents the number of publications on studies into dictionary use and language learning across study locations. It can be seen from the figure that countries or regions with more than five publications on this topic were Poland, China, USA, Taiwan, Japan, and Hong Kong, respectively. This finding confirms our general impression that some scholars from these countries or regions are rather active in the field of dictionary use and language learning, from example, Robert Lew and Anna Dziemianko from Poland, Yuzhen Chen and Hai Xu from China, and Alice Chan from Hong Kong. Also, it is interesting to note that scholars across all the continents publish studies on this topic. For instance, Dion Nkomo from South Africa in Africa, Jim Ranalli from the United States in North America, Vilson J. Leffa from Brazil in South America, and Helen Fraser from Australia in Oceania.

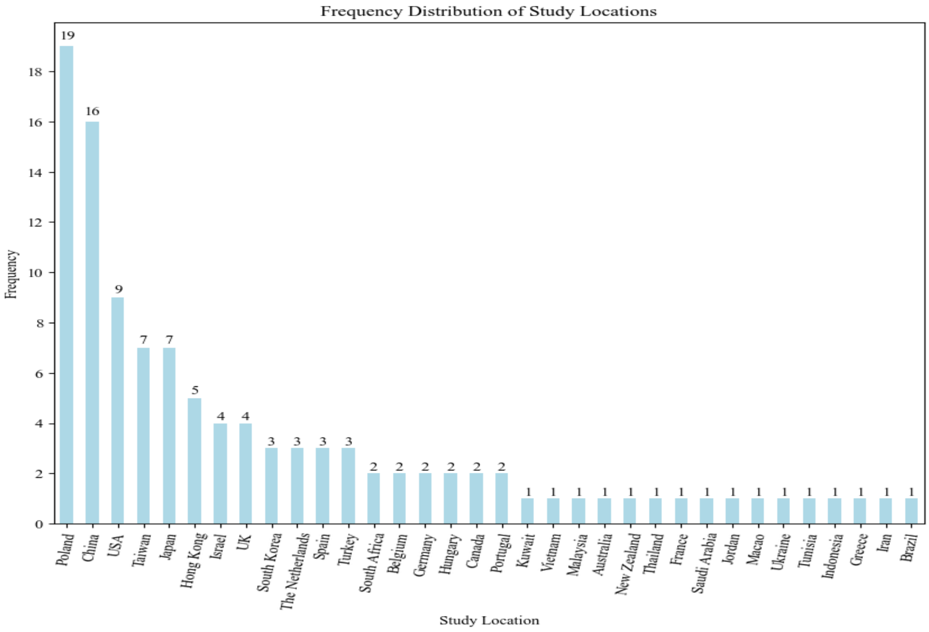


Figure 3: Distribution of studies across study location (Note that Taiwan, Hong Kong, and Macao are geographical concepts, and that they do not bear any political connotations)

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1940

4.3 Learners' L1s, L2s, proficiency levels, and educational levels involved in the selected studies

Figure 4 shows that participants in the included studies were from diverse L1 backgrounds, with a total of 47 languages involved. Out of these studies, 34 were conducted with Chinese as participants' L1, and 20 were conducted with Polish as the participants' L1, which echoes the finding that Poland and China ranked top two in terms of the number of publications.

Unlike participants' diverse L1s, their L2s involved in the selected studies were rather limited. Table 2 demonstrates the distribution of participants' L2s in the literature pertaining to dictionary use and language learning. It should be pointed out that although this scoping review was conducted mainly to chart the landscape of studies into dictionary use and L2 learning, two studies with native speakers as participants were also identified. Wolfer et al. (2018) conducted an empirical study to examine whether and to what extent lexicographical tools might help to improve L1 text revision results. Korat et al. (2022) compared whether e-book reading with a dictionary and the teacher's support, or e-book reading with a dictionary would outperform e-book reading only in fostering L1 kindergarten learners' vocabulary knowledge. These two studies, though limited in number, all consistently lent supporting evidence to the effectiveness of dictionary use in improving L1 learners' language learning outcomes.

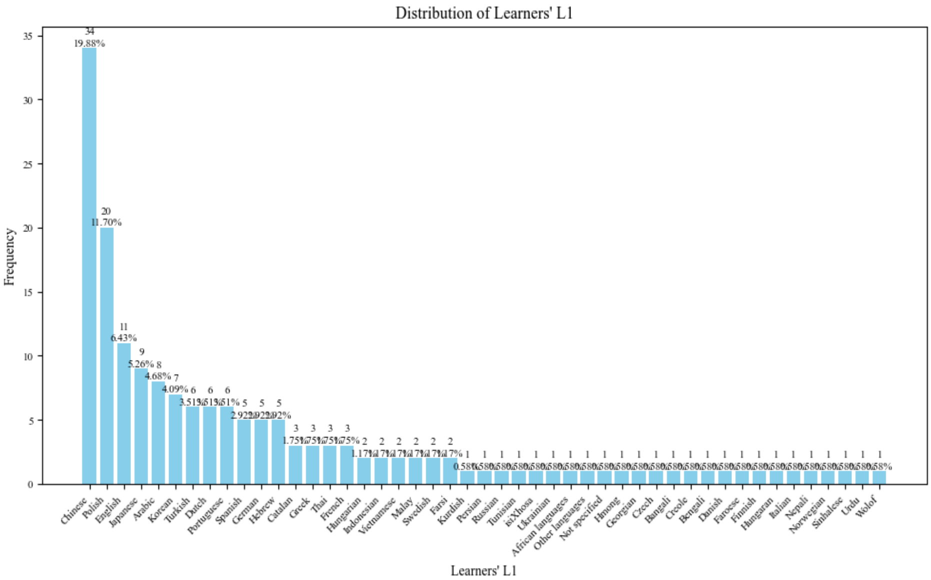


Figure 4: Learners' L1 backgrounds

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1940

Table 2: Learners' L2s involved in the studies

Target L2	N	Percentage
English	93	91.20%
French	3	2.94%
German	2	1.96%
Spanish	2	1.96%
Greek	1	0.98%
Chinese	1	0.98%

Out of the remaining 102 studies involving L2 learners, 93 were conducted with English as the L2, making it the most common target language being investigated among the included studies, followed by French (n = 3), German (n = 2), Spanish (n = 2), Greek (n = 1), and Chinese (n = 1), respectively. This result reinforced findings from Zhang et al. (2021) that the area of empirical dictionary use research has over-relied on conclusions drawn from L2 English learners, and that learners with other L2s, especially non-Indo-European languages, are under-represented. Such a result is probably not difficult to understand, as English has long been an international language and plays an important role in cross-national and cross-cultural exchanges, which has driven numerous L2 learners to devote themselves into English learning (McKay 2002). More studies are needed to investigate how different target L2s might influence the effectiveness of dictionary use on L2 learning outcomes, as language distance and script distance have been proved to exert an influence on L2 learning gains (Melby-Lervåg and Lervåg 2014).

Hulstijn and Atkins (1998) pointed out that users' l2 proficiency should be a main variable in empirical studies into dictionary use. Presented in Table 3 is the distribution of learners' L2 proficiency levels across studies. We treated each independent sample as an individual study, and therefore, there are 136 individual studies in total. It can be seen that one study did not report learners' proficiency. Out of the remaining studies, 54 dealt with advanced learners, 72 focused on intermediate learners, and 7 investigated low-level learners. In addition, 2 studies recruited participants of mixed proficiency levels. From the descriptive results, it can be concluded that almost all studies were conducted under instructional settings, as the proficiency levels were explicitly reported by the teacher researchers. Future studies should take account of learners using dictionaries in naturalistic settings, which, however, would increase the difficulty of data collection. Hulstijn and Atkins (1998) argued that participants of differed proficiency levels should be sampled when systematically investigating the effects of dictionary use on L2 learning outcomes. Our results revealed that a large majority of studies were conducted with advanced and intermediate learners as participants, and that only a small portion examined low-level learners. Therefore, low-level learners were under-represented in this area of research. This finding is consistent with Zhang et al. (2021), who also revealed in their

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1940

meta-analysis of studies into dictionary use and vocabulary learning that there is a scarcity of research that recruited low-level learners as participants. Also, it should be pointed out that a large proportion of studies relied on researchers' impersonal judgement or learners' grade levels in deciding their proficiency level, with only a handful of them categorizing learners' proficiency depending on scores from standardized tests such as TOFEL or IELTS. It remains unclear how such a hazard in proficiency categorization would influence the results. Future studies should make improvements in this regard by adopting standardized measures of proficiency when profiling participants.

Table 3: Learners' L2 proficiency levels

Proficiency	N	Percentage
Advanced	54	39.71%
Intermediate	72	52.94%
Low	7	5.15%
Mixed	2	1.47%
Not reported	1	0.74%

Table 4 demonstrates learners' educational levels. As is shown below, 86 studies were undertaken in higher or tertiary education context (e.g., Ptasznik 2020), followed by the 18 studies conducted in secondary education context (e.g., Bartelds 2021), and two in primary education context (Hall and Louw 2022; Tall and Hurman 2002). It is also worth noting that there was one study conducted in the preschool context (Korat et al. 2022). This finding is similar to scoping reviews conducted in other domains, such as digital game-based language learning (Hung et al. 2018), where researchers also found that college/university students were the most common research samples. This is probably because empirical dictionary use studies were mainly conducted by researchers from universities, where university students were more skilled dictionary users and it would be convenient to sample university students. Therefore, in order to draw a fuller picture of the effects of dictionary use and language learning, more studies are needed to investigate the dictionary lookup patterns of primary school students and kindergarteners.

Table 4: Learners' educational levels

Educational levels	N	Percentage
College	86	78.18%
Secondary school	18	16.36%
Mixed	3	2.73%
Primary school	2	1.82%
Kindergarten	1	0.91%

Taking together the results from Tables 3 and 4, it would be more obvious to see which groups of learners were relatively unrepresented. Table 5 is the cross tabulation showing the number of studies focusing on learners' proficiency levels by educational contexts. It can be seen that there is a severe scarcity of research dealing with kindergarten and primary school learners across all frequency levels.

Table 5: Number of studies focusing on learners' proficiency levels by educational contexts

Education levels	Language proficiency levels						Total
	Low	Intermediate	Advanced	Native	Mixed	Not specified	
Kindergarten	0	0	0	1	0	0	1
Primary	1	1	0	0	0	0	2
Secondary	0	8	2	0	2	0	12
Tertiary	2	33	21	1	23	1	81
Mixed	0	1	1	0	4	0	6
Other	0	0	1	0	1	0	2
Total	3	43	25	2	30	1	104

4.4 Dictionaries, dictionary type, and dictionary form

It is perhaps not surprising to find that among these included studies, the 'Big Six' dictionaries were frequently used by researchers in the field of dictionary use and language learning. The descriptive result shows that the most commonly used dictionary is the 'Oxford' series (n = 34), followed by the 'Longman' series (n = 30), the 'Collins' series (n = 14), the 'Cambridge' series (n = 12), the 'McMillan' series (n = 7), and the 'Merriam-Webster' series (n = 3), respectively. It mirrors the fact that researchers tend to appeal to authoritative dictionaries that have relatively large market share and learners are relatively familiar with when undertaking the studies.

In terms of dictionary type, 66 studies investigated L2 learners' use of monolingual dictionaries (e.g., Alzi'abi 2017; Chen 2022), 38 studies dealt with bilingual dictionaries (e.g., Chen and Liu 2023; Ptasznik 2023), and 17 focused on bilingualized dictionaries (e.g., Chan 2017; Kim 2018), respectively. In addition, there are four studies which did not specify which type of dictionary was adopted (e.g., Uchihara et al. 2022). These findings echo Welker's (2013) observation that "very few have investigated the use of monolingual dictionaries by native speakers," considering the observation was made a decade ago. Also, according to the meta-analysis conducted by Zhang et al. (2021), dictionary type was found to moderate the relationship between dictionary use and vocabulary learning. Therefore, we would suggest that future studies should explicitly specify the type of dictionary involved when investigating this topic so that results from different studies could be more interpretable and comparable.

As for dictionary form, it is interesting to see that electronic dictionaries play a dominant role in the included studies (n = 53), accounting for 50.96%. There were 38 studies were conducted based on paper dictionaries, taking up 36.54% of the total. Twelve studies used both electronic and paper dictionaries (e.g., Alahmadi and Foltz 2020), accounting for 11.54%. One study failed to specify which dictionary form was chosen (Fajt et al. 2023) quite possibly due to the fact that the study was designed as a large-scale survey to investigate the relationship between learners' motivation and dictionary use. The dominance of electronic dictionaries in such studies reflect the impact of the Digital Revolution on dictionary use research, as mentioned earlier in Section 4.2, where researchers have begun to shift their attention onto electronic dictionaries (Müller-Spitzer 2014).

4.5 Study focus and language learning outcomes

The term "study focus" refers to the specific linguistic aspects that the research within the domain of dictionary use and language learning addresses, including vocabulary, writing, reading, grammar, translation, and pronunciation. Reflecting the foundational importance of vocabulary in language acquisition, it is anticipated that a substantial portion of these studies would concentrate on this area. This expectation is confirmed by the data presented in Table 6, where more than half of the studies (n=68; 52%) have a primary focus on vocabulary enhancement.

As detailed in Table 6, writing and reading are the subsequent areas of focus, with 18% (n=23) and 15% (n=20) of the studies dedicated to each, respectively. Studies focusing on grammar and translation are less prevalent, accounting for 6% (n=8) and 5% (n=6) of the total, respectively. Additionally, a small subset of studies (n=4) defies straightforward categorization into these linguistic aspects. Among these, two studies (Fajt et al. 2023; Liu et al. 2019) examine the motivations behind dictionary use, one (Ptasznik 2020) explores defining models, and another (Nkomo 2017) investigates dictionary use behavior during examinations. Notably, only one study (Fraser 1997), dating back nearly three decades, has specifically addressed the aspect of pronunciation.

Given the potential of electronic dictionaries and dictionary applications in assisting language learners with pronunciation, the relative neglect of this area in recent research is striking. This observation, highlighted in Table 6, suggests a gap that warrants further investigation. Recent survey studies (El-Sawy 2019; Hakim et al. 2020; Metruk 2017) indicate that a significant portion of EFL learners worldwide use the pronunciation features of electronic dictionaries to improve their own pronunciation, yet relevant experimental studies examining the extent to which electronic dictionaries could influence L2 pronunciation learning as well as the potential moderator variables remain insufficient. Future studies should be conducted to explore the use of these technological resources in the context of pronunciation learning, thereby contributing to a more comprehen-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1940

sive understanding of how dictionaries can be leveraged to support language acquisition holistically.

Table 6: Distribution of study focus

Study Focus	**Frequency**	**Percentage**
Vocabulary	68	52.30%
Writing	23	17.69%
Reading	20	15.38%
Grammar	8	6.15%
Translation	6	4.62%
Other	4	3.08%
Pronunciation	1	0.77%

5. Conclusion

This scoping review assembles a large body of empirical studies examining the effects of dictionary use on various aspects of language learning. It illuminates the evolving landscape of dictionary use and language learning research, highlighting both its strengths and areas for further exploration. The dominance of experimental approaches and the increasing adoption of mixed-method designs reflect the field's maturation and commitment to rigorous research practices. However, this review also reveals several gaps and challenges that warrant further attention. While the rapid development of electronic dictionaries and dictionary applications offers new opportunities for language learning research, studies into their potential as a learning tool are limited and confined to their impact on the improvement of vocabulary knowledge, reading comprehension, and writing competence. Further efforts are needed to explore the effectiveness of these technological resources in supporting language learning in areas like grammar, pronunciation, and pragmatic competence. Also, although researchers have begun to focus their attention on learners' dictionary look-up behavior and their cognitive processes involved in it, much of the data were collected using questionnaires or surveys. Researchers are encouraged to explore methods such as screen recordings, eye-tracking and think-aloud protocols to gain deeper insights into how learners interact with dictionaries in authentic contexts. In addition, it is noted that a large proportion of the sampled studies adopted a cross-sectional study, which limits our understanding of the long-term effects of dictionary use on language learning. Given that longitudinal studies boast higher ecological validity, studies with longitudinal design are crucially needed to track the changes of learners' dictionary use skills and reveal their sustained impact on language proficiency over time. Finally, the over-reliance on English as a second language and the underrepresentation of low-level and primary/secondary school learners necessitate a more inclusive research agenda. Future studies should investigate the unique challenges and opportunities faced by diverse language learner populations and explore the potential of dictionaries in supporting their language learning processes. By addressing these gaps and challenges, we believe that researchers could contribute to a more thorough understanding of the interplay between dictionary use and language learning, which will facilitate the design of more user-friendly dictionaries, formulate pedagogical strategies that cater to the diverse needs of language learners, and harness the full potential of dictionaries as powerful tools for language learning.

Admittedly however, it is important to acknowledge the limitations of the review when interpreting these findings. The synthesis of studies is confined to the search terms used and the journals examined, potentially overlooking relevant monographs, conference papers and book chapters. Therefore, it is recommended that future reviews should aim for a broader scope, covering a wider array of journals, books, and conference proceedings. Additionally, it should be noted that we only included literature written in English, which further limits the scope of the study.

To recapitulate, this study should be recognized as a pioneering effort in systematically synthesizing research on dictionary use and language learning. To our knowledge, this is the first study to introduce the scoping review methodology to the field of lexicography, and it provides a guide for researchers in this area as to how a scoping review could be conducted. Also, by systematically examining the research methods applied, the demographics of the learners involved, the dictionaries used, and the research foci of the sampled studies, this study not only map out the current landscape of research in this field but also serve as a guidepost for identifying research gaps and charting the course for future scholarly endeavors.

Acknowledgements

This work was supported by the Guangdong Planning Office of Philosophy and Social Science Fund (Grant number: GD23WZXY02-05), the Humanities and Social Science Fund of Ministry of Education of the People's Republic of China (Grant number: 23YJC740090), the National Office of Philosophy and Social Sciences Fund (Grant number: 22BYY010), and the National Office of Philosophy and Social Sciences Fund (Grant number: 20AYY012).

References

Alahmadi, A. and A. Foltz. 2020. Exploring the Effect of Lexical Inferencing and Dictionary Consultation on Undergraduate EFL Students' Vocabulary Acquisition. *PLOS ONE* 15(7): 1-25.

Alzi'abi, S.E. 2017. Guessing Verb–Adverb Collocations: Arab EFL Learners' Use of Electronic Dictionaries. *Lexikos* 27: 50-77.

Arksey, H. and L. O'Malley. 2005. Scoping Studies: Towards a Methodological Framework. *International Journal of Social Research Methodology* 8(1): 19-32.

Atkins, B.T.S. (Ed.). 1998. *Using Dictionaries: Studies of Dictionary Use by Language Learners and Translators.* Tübingen: Max Niemeyer.

Bartelds, D. 2021. How To Stay in the Loop. A Think-Aloud Study on Dictionary Use by Excellent Secondary-School Students of Ancient Greek. *International Journal of Lexicography* 34: 453-471.

Chan, A.Y.W. 2017. The Effectiveness of Using a Bilingualized Dictionary for Determining Noun Countability and Article Selection. *Lexikos* 27: 183-213.

Chen, Y. 2022. The Effect of Learning Conditions on Collocation Gains: A Case Study of Task-based Dictionary Use Instruction. *Lexikos* 32: 1-30.

Chen, Y. and S. Liu. 2022. Exploring the Use of an Online Bilingual Dictionary in EFL Writing. *International Journal of Lexicography* 35(4): 468-490.

Chen, Y. and S. Liu. 2023. A Further Look into the Use of a Dictionary APP in EFL Writing: A Replication Study. *Lexikos* 33: 324-349.

Doležal, F.T. and D.R. McCreary. 1999. *Pedagogical Lexicography Today : A Critical Bibliography on Learners' Dictionaries with Special Emphasis on Language Learners and Dictionary Users.* Tübingen: Max Niemeyer.

Dziemianko, A. 2022. The Usefulness of Graphic Illustrations in Online Dictionaries. *ReCALL* 34(2): 218-34.

El-Sawy, H.E.A. 2019. Electronic and Student-created Dictionaries for Enhancing EFL Pronunciation and Vocabulary Usage. *Theory and Practice in Language Studies* 9(9): 1088-1099.

Fajt, B., M. Bánhegyi and K.P. Márkus. 2023. The Interrelationship between EFL Learning Motivation and Dictionary Use. *International Journal of Lexicography* 37(1): 1-16.

Fellbaum, C. 2014. Large-Scale Lexicography in the Digital Age. *International Journal of Lexicography* 27(4): 378-395.

Fraser, H. 1997. Dictionary Pronunciation Guides for English. *International Journal of Lexicography* 10(3): 181-208.

Hakim, M.A.R., S.N. Aryati and D. Kurniawan. 2020. Investigating E-Dictionaries on Speaking Ability among University Students in Malaysia. *Universal Journal of Educational Research* 8(12): 6536-6551.

Hall, M. and P. Louw. 2022. The Perceived Impacts of a Bilingual Learner's Dictionary. *International Journal of Lexicography* 35(3): 273-295.

Hartmann, R.R.K. 2001. *Teaching and Researching Lexicography*. London: Routledge.

Hulstijn, J.H. and B.T.S. Atkins. 1998. Empirical Research on Dictionary Use in Foreign-Language Learning: Survey and Discussion. Atkins, B.T.S. (Ed.). 1998. *Using Dictionaries: Studies of Dictionary Use by Language Learners and Translators:* 7-19. Tübingen: Max Niemeyer.

Hung, H.-T., J.C. Yang, G.-J. Hwang, H.-C. Chu and C.-C. Wang. 2018. A Scoping Review of Research on Digital Game-based Language Learning. *Computers & Education* 126: 89-104.

Kim, S. 2018. A Lexicographic Approach to Teaching the English Article System: Help or Hindrance? *Lexikos* 28: 196-220.

Korat, O., S. Atishkin and O. Segal-Drori. 2022. Vocabulary Enrichment Using an E-Book with and without Kindergarten Teacher's Support among LSES Children. *Early Child Development and Care* 192(9): 1384-1401.

Krippendorff, K. 2018. *Content Analysis: An Introduction to Its Methodology.* Los Angeles: SAGE.

Lew, R. 2007. Herbert Andreas Welker. *O Uso de Dicionários: Panorama Geral Das Pesquisas Empíricas. International Journal of Lexicography* 20(4): 401-403.

Lew, R. 2011. Studies in Dictionary Use: Recent Developments. *International Journal of Lexicography* 24(1): 1-4.

Lew, R. and G.-M. de Schryver. 2014. Dictionary Users in the Digital Revolution. *International Journal of Lexicography* 27(4): 341-359.

L'Homme, M.-C. and M.C. Cormier. 2014. Dictionaries and the Digital Revolution: A Focus on Users and Lexical Databases. *International Journal of Lexicography* 27(4): 331-340.

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.

Liu, X., D. Zheng and Y. Chen. 2019. Latent Classes of Smartphone Dictionary Users among Chinese EFL Learners: A Mixed-method Inquiry into Motivation for Mobile Assisted Language Learning. *International Journal of Lexicography* 32(1): 68-91.

Lo, S. 2024. The Effects of NMT as a de facto Dictionary on Vocabulary Learning: A Comparison of Three Look-up Conditions. *Computer Assisted Language Learning*: 1-21.

Mak, S. and A. Thomas. 2022. Steps for Conducting a Scoping Review. *Journal of Graduate Medical Education* 14(5): 565-567.

McCreary, D.R. and F. Doležal. 1998. Language Learners and Dictionary Users: Bibliographic Findings and Commentary. Fontenelle, Thierry, Philippe Hiligsmann, Archibald Michiels, André Moulin and Siegfried Theissen (Eds.). 1998. *Proceedings of the Eighth EURALEX International Congress, EURALEX 1998, Liège, Belgium, August 4–8, 1998:* 611-618. Liège: English and Dutch Departments, University of Liège.

McKay, S.L. 2002. *Teaching English as an International Language : Rethinking Goals and Approaches.* Oxford: Oxford University Press.

Melby-Lervåg, M. and A. Lervåg. 2014. Reading Comprehension and Its Underlying Components in Second-Language Learners: A Meta-Analysis of Studies Comparing First- and Second-Language Learners. *Psychological Bulletin* 140(2): 409-433.

Metruk, R. 2017. The Use of Electronic Dictionaries for Pronunciation Practice by University EFL Students. *Teaching English with Technology* 17(4): 38-51.

Müller-Spitzer, C. (Ed.). 2014. *Using Online Dictionaries*. Berlin/Boston: De Gruyter.

Nesi, H. 2014. Dictionary Use by English Language Learners. *Language Teaching* 47(1): 38-55.

Nkomo, D. 2017. The Dictionary in Examinations at a South African University: A Linguistic or a Pedagogic Intervention? *Lexikos* 27: 346-377.

O'Flaherty, J. and C. Phillips. 2015. The Use of Flipped Classrooms in Higher Education: A Scoping Review. *Internet and Higher Education* 25: 85-95.

Ptasznik, B. 2020. Which Defining Model Contributes to More Successful Extraction of Syntactic Class Information and Translation Accuracy? *Lexikos* 30: 363-385.

Ptasznik, B. 2023. More Examples May Benefit Dictionary Users. *International Journal of Lexicography* 36(1): 29-55.

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.

Ronald, J. 2003a. A Review of Research into Vocabulary Acquisition through Dictionary Use . Part 1: Intentional Vocabulary Learning through Dictionary Use. *Studies in the Humanities and Sciences* 44(1): 285-307.

Ronald, J. 2003b. A Review of Research into Vocabulary Acquisition through Dictionary Use. Part 2 : Incidental Vocabulary Acquisition through Dictionary Use. *Studies in the Humanities and Sciences* 44(2): 67-97.

Sinclair, J.M. 1984. Lexicography as an Academic Subject. Hartmann, Reinhard R.K. (Ed.). 1984. *LEXeter '83: Proceedings. Papers from the International Conference on Lexicography at Exeter, England, 9–12 September 1983:* 3-12. Tübingen: Max Niemeyer.

Stockwell, G. 2007. A Review of Technology Choice for Teaching Language Skills and Areas in the CALL Literature. *ReCALL* 19(2): 105-120.

Tall, G. and J. Hurman. 2002. Using Dictionaries in Modern Language GCSE Examinations. *Educational Review* 54(3): 205-217.

Tarp, S. 2009. Reflections on Lexicographical User Research. *Lexikos* 19: 275-296.

Töpel, A. 2014. Review of Research into the Use of Electronic Dictionaries. Müller-Spitzer, C. (Ed.). 2014. *Using Online Dictionaries:* 13-54. Berlin/Boston: De Gruyter.

Tsai, K.-J. 2019. Corpora and Dictionaries as Learning Aids: Inductive versus Deductive Approaches to Constructing Vocabulary Knowledge. *Computer Assisted Language Learning* 32(8): 805-826.

Uchihara, T., M. Eguchi and J. Clenton. 2022. The Contribution of Guessing from Context and Dictionary Use to Receptive and Productive Vocabulary Knowledge: A Structural Equation Modeling Approach. *Language Teaching Research*: 1-24.

Welker, H.A. 2010. *Dictionary Use: A General Survey of Empirical Studies.* Brasília: Author's Edition.

Welker, H.A. 2013. Empirical Research into Dictionary Use since 1990. Gouws, R.H., U. Heid, W. Schweickard and H.E. Wiegand (Eds.). 2013. *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography:* 531-540. Berlin/Boston: De Gruyter Mouton.

Wolfer, S., T. Bartz, T. Weber, A. Abel, C.M. Meyer, C. Müller-Spitzer and A. Storrer. 2018. The Effectiveness of Lexicographic Tools for Optimising Written L1-Texts. *International Journal of Lexicography* 31(1): 1-28.

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.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1941 (Article)

Aspekte van inligtingsonttrekkingstrukture 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)
en
Theo J.D. Bothma, *Departement Inligtingkunde,*
Universiteit van Pretoria, Suid-Afrika
(theo.bothma@up.ac.za) (https://orcid.org/0000-0001-7850-3263)

Opsomming: In gedrukte woordeboeke met 'n statiese artikel- en boekstruktuur is die verhouding tussen die verspreiding van data en die onttrekking van inligting eenvoudig en die meeste woordeboeke volg min of meer dieselfde ordeningstelsels. Aanlyn woordeboeke vertoon dinamiese strukture met data wat dikwels op verskillende vlakke en in verskillende vakke aangebied word. Dit dwing die gebruiker om afwaarts (of vertikaal) en sywaarts (of horisontaal) in 'n artikel of bepaalde artikelkommentare te beweeg en om soekprosedures te volg wat nuwe strukture soos uitgebreide kommentare en selfs nuwe soekposisies soos soektonnels in ag moet neem. Dit is uiters belangrik dat gebruikers bewus moet wees van die volle spektrum data in 'n leksikografiese produk en dat hulle daartoe in staat moet wees om die datatipe wat vir hulle ter sake is, te kan kies. Gebruikers van leksikografiese produkte het ondersteuning nodig om 'n optimale onttrekking van inligting te verseker. Gevolglik moet leksikograwe moeite doen om 'n inligtingsonttrekkingstruktuur te skep wat gebruikers kan help om suksesvolle woordeboekraadpleging te verseker. Dit kan verfyning van die databasisstruktuur van die woordeboek vereis, en die ontwikkeling van 'n gevorderde koppelvlakontwerp wat nuwe kenmerke en navigasie-opsies aan die gebruiker beskikbaar kan stel. Dit beklemtoon weer eens die noodsaaklikheid van 'n multidissiplinêre span in die ontwikkeling van 'n woordeboek om 'n geskikte finale produk aan die gebruiker te kan bied. Hierdie artikel bied voorstelle vir gestruktureerde en kitstoegang tot data en die gebruik van datamerkers om gebruikers te lei na aanduiders in komplekse woordeboekartikels. Die klem is ook op die gebruikersvriendelikheid van leksikografiese instrumente wat die gehalte van intuïtiewe woordeboekgebruik kan verbeter.

Sleutelwoorde: AANLYN WOORDEBOEKE, ARTIKELSTRUKTUUR, BEPERKTE ARTIKEL, INLIGTINGSONTTREKKINGSTRUKTUUR, KONTEKSTUALISERING, OMVATTENDE ARTIKEL, SEMANTIESE SUBKOMMENTAAR, SKAKELING, SOEKSONE, SOEKTONNEL, VAKKE, VLAKKE

Abstract: Aspects of information retrieval structures in online dictionaries. In printed dictionaries with a static article and book structure the relation between inserting data and retrieving information is not complex, and most dictionaries follow more or less the same ordering systems. Online dictionaries display dynamic structures with data often presented in

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1941 (Article)

various levels and compartments that compel the user to drill down (or vertically) into articles or laterally (or horizontally) within specific comments of articles, and to embark on search procedures that must negotiate new structures such as extended comments and even new search positions such as search tunnels. It is of paramount importance that users should be aware of the full spectrum of data included in a lexicographic product and should be able to select the data types of interest to them. Users of lexicographic tools need assistance to ensure an optimal retrieval of information from the data. Consequently, lexicographers need to embark on dedicated ways to establish an information retrieval structure that can enable users to achieve an adequate dictionary consultation. This might demand a refinement of the database of the dictionary as well as a sophisticated interface that can provide new features and navigation options to the user. It emphasizes yet again the need for a multidisciplinary team in the development of a dictionary to provide the user with an appropriate final product. This paper offers proposals for structured and rapid access to data and the use of data indicators to guide users to items occurring in complex dictionary articles. The emphasis is also on user-friendliness of lexicographic instruments that can enhance the quality of intuitive dictionary use.

Keywords: ARTICLE STRUCTURE, COMPARTMENTS, COMPREHENSIVE ARTICLE, CONTEXTUALISATION, INFORMATION RETRIEVAL STRUCTURE, LEVELS, LINKING, ONLINE DICTIONARIES, RESTRICTED ARTICLE, SEARCH TUNNEL, SEARCH ZONE, SUBCOMMENT ON SEMANTICS

1. Inleiding en agtergrond

Hartmann en James (1998) bied die volgende definisie vir die lemma *lexicography*:

> The professional activity and academic field concerned with DICTIONARIES and other REFERENCE WORKS. It has two basic divisions: lexicographic practice, or DICTIONARY-MAKING, and lexicographic theory, or DICTIONARY RESEARCH. ...

In hierdie bydrae gaan dit primêr om inligtingsonttrekkingstrukture in aanlyn woordeboeke; tog sal ander leksikografiese naslaanbronne wat nie woordeboeke is nie, ook ter sprake kom. Daarom is Hartmann en James se verwysing na "DICTIONARIES and other REFERENCE WORKS" ter sake.

'n Gereelde verwysing na Samuel Johnson se woordeboek van 1755 is na sy definisie van 'n leksikograaf as "A writer of dictionaries; a harmless drudge, ..." Daar word selde ook aandag gegee aan die daaropvolgende deel van die betekenisparafrase, naamlik "… that busies himself in tracing the original, and detailing the signification of words." Die aktiwiteit van besig wees met betekenisverklaring vorm 'n kerndeel van 'n leksikograaf se opdrag. In hierdie artikel sal dit juis die onttrekking van semantiese data en meer spesifiek die betekenisparafrase wees waaraan daar aandag gegee word. Dit word gedoen met die wete dat die leksikografie oor 'n wye tipologiese verskeidenheid beskik en dat hierdie woordeboeke 'n wye verskeidenheid datatipes bevat wat ook herwinbaar moet wees.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1941 (Article)

Met die aanvang van die beplanning en samestelling van 'n woordeboek moet die leksikograaf duidelikheid hê oor wie die veronderstelde teikengebruikers van die woordeboek gaan wees, asook wat hulle leksikografiese behoeftes en naslaanvaardighede is. Vergelyk hier onder andere Hartmann (1989), Tarp (2008) en Wiegand (1998). Op grond van die teikengebruiker en hulle behoeftes besluit die leksikograaf wat die leksikografiese funksies van die woordeboek gaan wees. Die leksikografiese funksies is dan van bepalende belang vir die keuse van die leksikografiese data en die woordeboekstrukture wat nodig is vir die aanbieding en huisvesting van hierdie data. 'n Wesenlike deel van die leksikograaf se beplanning is weliswaar gerig op die keuse maar ook op die aanbieding van data ter bevrediging van die gebruiker se leksikografiese behoeftes. Die uiteindelike sukses van 'n leksikograaf se werk word bepaal deur die sukses van die teikengebruikers wanneer hulle die woordeboek raadpleeg. Om hierdie sukses te verseker, is dit belangrik dat woordeboekstrukture ontwerp en aangewend moet word om toegang tot die leksikografiese data te verhoog sodat die gepaste leksikografiese funksie bevredig kan word.

Die leksikograaf mag die suksesvolle onttrekking van inligting deur die gebruiker nie as 'n vanselfsprekendheid beskou nie. Ook hier moet 'n struktuur beplan en aangewend word om sukses te verseker. 'n Gepaste inligtingsonttrekkingstruktuur is noodsaaklik om woordeboekgebruik te kan optimaliseer. Aspekte van hierdie struktuurtipe wat nog slegs beperkte aandag in die metaleksikografiese literatuur gekry het, word in hierdie artikel bespreek. Woordeboekstrukture tree nie geïsoleerd van mekaar op nie en in 'n bespreking van die inligtingsonttrekkingstruktuur moet ander strukture, onder meer die dataverspreidingstruktuur en die artikelstruktuur, ook aan die orde gestel word.

2. Die dataverspreidingstruktuur

Die dataverspreidingstruktuur, vergelyk Bergenholtz et al. (1999), bepaal die toewysing van leksikografiese data aan spesifieke soekposisies in woordeboeke. Dié struktuur word aangewend in sowel gedrukte as aanlyn woordeboeke, asook binne 'n breër leksikografiese soekomgewing. Verskillende soekposisies kan onderskei word. Wiegand et al. (2013: 63) onderskei tussen 'n soeksone, 'n soekgebied en 'n soekveld. Die soeksone is die gleuf in 'n woordeboekartikel waarin 'n aanduider geplaas word. Elke woordeboekartikel is 'n soekgebied en die sentrale teks van 'n woordeboek wat al die artikeltrajekte bevat, is 'n soekveld. Daarbenewens maak Gouws (2023: 383-384) ook voorsiening vir 'n soekstreek, dit is die volle woordeboek met sy sentrale teks plus die verskillende buitetekste; 'n soekdomein, dit is die portaal waarbinne aanlyn woordeboeke geplaas kan word; en 'n soekuniversum, dit is die omgewing buite die soekdomein, byvoorbeeld die internet of ander woordeboekportaal-eksterne bronne waartoe gebruikers via die woordeboek toegang het.

Dit is belangrik om daarop te let dat 'n inligtingsonttrekkingstruktuur so geformuleer moet word dat dit die gebruiker kan help om leksikografiese inlig-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1941 (Article)

ting aan die data in al die verskillende soekposisies te kan onttrek. In hierdie artikel is die fokus beperk tot die woordeboekartikel as soekgebied en die onttrekking van inligting aan data in bepaalde soeksones.

3. Die soekgebied

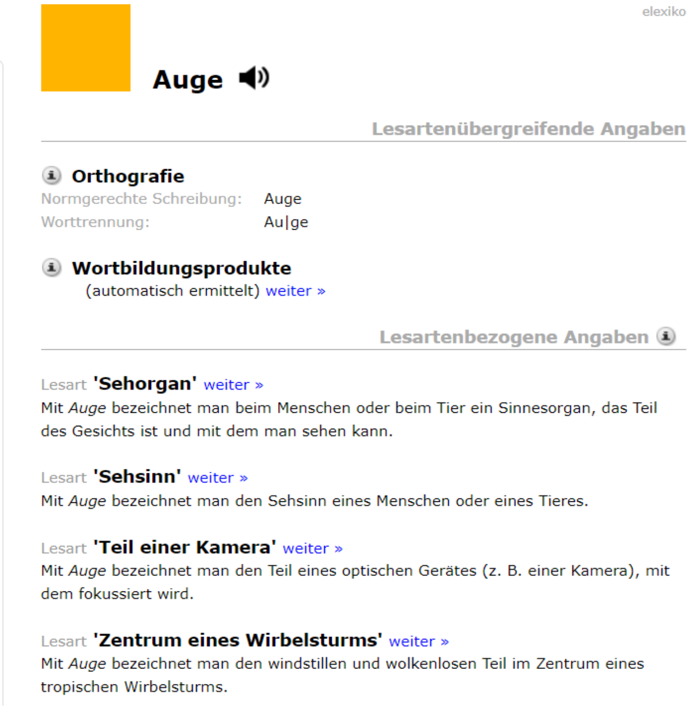
Die woordeboekartikel is die primêre posisie vir die toesegging van leksikografiese data in 'n woordeboek. In teenstelling met die statiese aard van woordeboekartikels in gedrukte woordeboeke het aanlyn woordeboeke 'n dinamiese artikelstruktuur (Gouws 2014: 164). Aanlyn woordeboeke mag dalk wel haas onbeperkte bergingsruimte hê, maar omdat daar beperkte vertoonruimte is, kan al die data wat aan 'n artikel toegewys word, nie noodwendig tegelykertyd gesien word nie. Aanlyn woordeboeke het dikwels omvattende artikels wat verskeie beperkte artikels insluit (Gouws 2014). Beperkte artikels is onderafdelings van 'n omvattende artikel en elkeen bied slegs 'n deel van die leksikografiese bewerking, byvoorbeeld een of meer soeksones. Die omvattende artikel kan nie as 'n geheel gesien word nie en toegang tot die data geskied via die verskillende beperkte artikels.

Die verspreiding van data in 'n woordeboekartikel verseker nie gebruikersukses nie. Wat die leksikograaf in 'n omvattende artikel plaas, moet die gebruiker kan vind en onttrek. Daarvoor is die bykomende strukture nodig.

4. Artikelstruktuur: kommentare, vlakke en vakke

Volgens Wiegand (1989: 470) bestaan 'n woordeboekartikel met 'n basiese struktuur uit twee kommentare, naamlik 'n vormkommentaar en 'n semantiese kommentaar. Alhoewel die artikelstruktuur van aanlyn woordeboeke nie noodwendig dieselfde byeenplasing van aanduiders in die onderskeie komponente vertoon nie, kan 'n vergelykbare indeling dikwels steeds herken word (Gouws 2014). Soos in gedrukte woordeboeke bevat die semantiese kommentaar van artikels met 'n lemma wat 'n polisemiese leksikale item verteenwoordig, dikwels ook verskillende semantiese subkommentare wat elk huisvesting aan die bewerking van een polisemiese onderskeiding bied. Die fokus van hierdie bespreking is vervolgens op die onttrekking van inligting aan sulke semantiese subkommentare. Dit word gedoen aan die hand van artikels uit een aanlyn woordeboek, naamlik die Leibniz-Institut für Deutsche Sprache (IDS) se *elexiko* wat op OWID se woordeboekportaal verskyn.

In 'n omvattende woordeboekartikel wat verskillende semantiese subkommentare vertoon, is daar 'n vertikale hiërargiese ordening van hierdie semantiese subkommentare en wel so dat elke subkommentaar as 'n afsonderlike vlak van die semantiese kommentaar optree en 'n eie beperkte artikel daarstel. Vlakke kan hier beskou word as soekposisies binne 'n kommentaar van 'n omvattende artikel wat elk beset word deur 'n beperkte artikel wat een semantiese subkommentaar bevat. Opeenvolgende vlakke kom voor in struktureel laer artikelposisies en die verskillende vlakke se data is nie gelyktydig sigbaar nie. Toegang tot die data in 'n bepaalde vlak vereis dat daar vanuit die betrokke kommentaar se openingsaanbod afwaarts beweeg moet word. Vergelyk in hierdie verband figuur 1, 'n deelartikel van *elexiko* se lemma *Auge* (= "oog"). In hierdie beperkte artikel, die openingskermskoot wanneer die lemma *Auge* gesoek word, is die artikelkomponent "Lesartenübergreifende Angaben" (= betekenisonderskeidingoorkoepelende aanduiders") vergelykbaar met die semantiese kommentaar van 'n artikel. Dit bevat 'n lysting van die verskillende betekenisonderskeidinge van die woord *Auge* met 'n eenwoordbetekenissamevatting en 'n bondige betekenisparafrase:



elexiko

Auge

Lesartenübergreifende Angaben

Orthografie
Normgerechte Schreibung: Auge
Worttrennung: Au|ge

Wortbildungsprodukte
(automatisch ermittelt) weiter »

Lesartenbezogene Angaben

Lesart **'Sehorgan'** weiter »
Mit *Auge* bezeichnet man beim Menschen oder beim Tier ein Sinnesorgan, das Teil des Gesichts ist und mit dem man sehen kann.

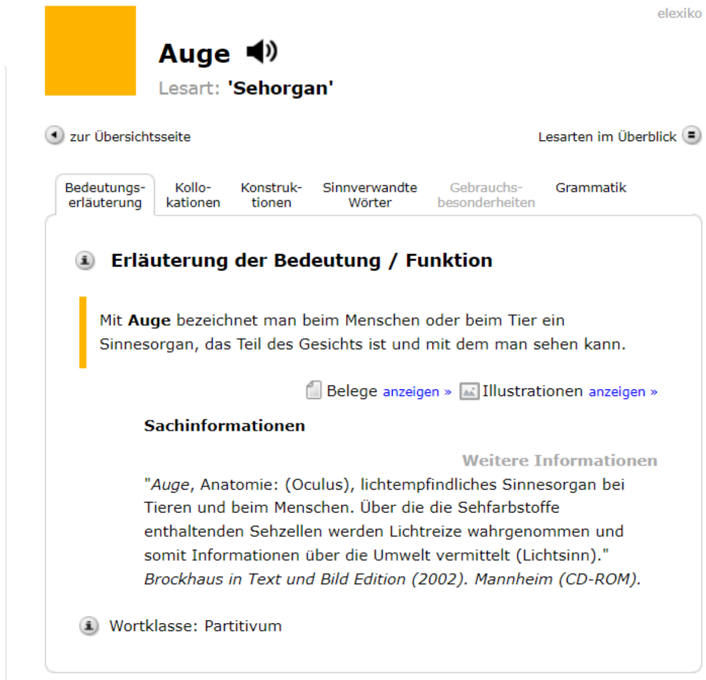
Lesart **'Sehsinn'** weiter »
Mit *Auge* bezeichnet man den Sehsinn eines Menschen oder eines Tieres.

Lesart **'Teil einer Kamera'** weiter »
Mit *Auge* bezeichnet man den Teil eines optischen Gerätes (z. B. einer Kamera), mit dem fokussiert wird.

Lesart **'Zentrum eines Wirbelsturms'** weiter »
Mit *Auge* bezeichnet man den windstillen und wolkenlosen Teil im Zentrum eines tropischen Wirbelsturms.

Figuur 1: Uit *elexiko*

Elke eenwoordbetekenissamevatting word gevolg deur 'n struktuurmerker *weiter* (= "verder"). Dit is 'n skakel na 'n volgende vlak van die semantiese kommentaar en lei die gebruiker na die betrokke beperkte artikel waarin meer data oor die tersaaklike betekenisonderskeiding verstrek word. 'n Klik op die skakel *weiter* ten opsigte van die eerste betekenisonderskeiding *Sehorgan* (= "sigorgaan") van *Auge* lei tot 'n afwaartse beweging na 'n vlak wat die beperkte artikel bevat wat in figuur 2 gesien kan word:



Auge

Lesart: **'Sehorgan'**

Erläuterung der Bedeutung / Funktion

Mit **Auge** bezeichnet man beim Menschen oder beim Tier ein Sinnesorgan, das Teil des Gesichts ist und mit dem man sehen kann.

Sachinformationen

"*Auge*, Anatomie: (Oculus), lichtempfindliches Sinnesorgan bei Tieren und beim Menschen. Über die die Sehfarbstoffe enthaltenden Sehzellen werden Lichtreize wahrgenommen und somit Informationen über die Umwelt vermittelt (Lichtsinn)." *Brockhaus in Text und Bild Edition (2002). Mannheim (CD-ROM).*

Wortklasse: Partitivum

Figuur 2: Uit *elexiko*

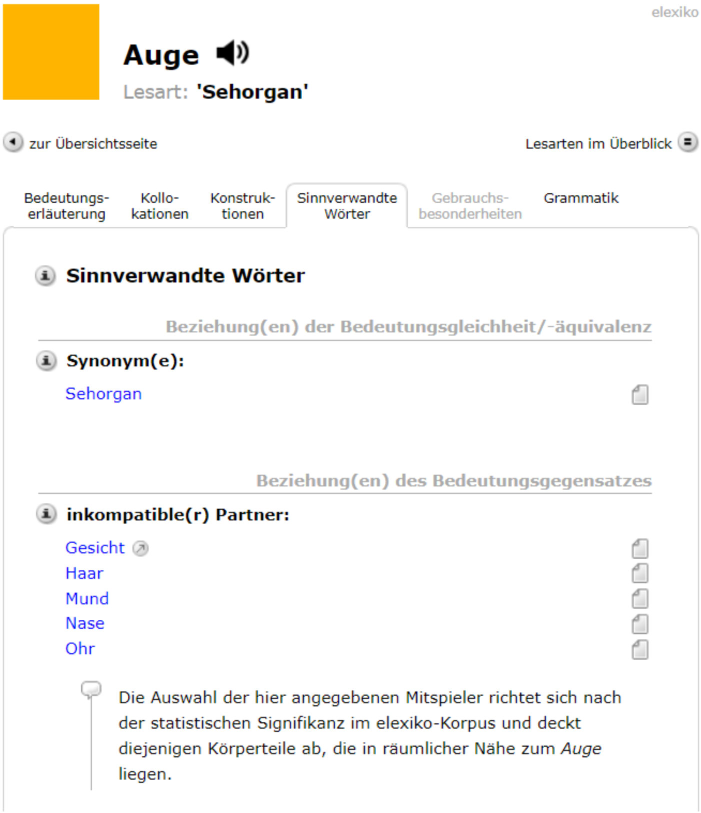
Hierdie beperkte artikel het ook *Auge* as lemma maar hier gaan dit nie om die polisemiese leksikale item *Auge* nie, maar slegs om die lemma as gidselement vir 'n bewerking van een van die leksikale item se betekenisonderskeidings. Al die data wat in hierdie beperkte artikel voorkom, is slegs op *Auge* se optrede in een betekenisonderskeiding gerig. Die verskillende datatipes wat in een vlak voorkom en betrekking het op die enkele betekenisonderskeiding, word ook nie gelyktydig vertoon nie, maar kan wel bereik word deur binne die betrokke vlak rond te beweeg. Hierdie betekenisonderskeiding-spesifieke data word in verskillende vakke in die vlak verpak. Vakke kan hier beskou word as soekposisies binne 'n subkommentaar, 'n vlak, van 'n omvattende artikel wat elk beset word deur 'n beperkte artikel wat een datatipe bevat wat spesifiek gerig is op die betrokke subkommentaar.

Naas ander inskrywings bevat die beperkte artikel in figuur 2 ook verskeie data-merkers (*Kollokation*, *Konstruktion*, ensovoorts) wat op die boonste dwarslyn verstrek word. 'n Klik op enigeen van hierdie merkers lei die gebruiker na 'n vak binne dieselfde vlak waarin bykomende data verstrek word wat nie die lemma van die omvattende artikel as adres het nie, maar wel die lemma van die betrokke beperkte artikel. Dit is bykomende data ter sake vir die optrede van die leksikale item in daardie spesifieke betekenisonderskeiding. Om die data in

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1941 (Article)

'n vak te bereik, vereis nie 'n afwaartse soekbeweging na 'n volgende vlak nie, maar slegs 'n horisontale (of sywaartse) beweging binne die betrokke vlak.

Die gebruik van verskillende vakke in een vlak help die leksikograaf om data-oorlading in die aanbieding van 'n bepaalde semantiese subkommentaar te voorkom. Dit verhoog ook die gebruikersvriendelikheid van die woordeboek omdat gebruikers slegs hoef te klik op 'n merker wat relevant is vir hulle soektog en hulle ontsluit sodoende slegs die vak met data wat hulle benodig. Om hierdie funksionaliteite moontlik te maak, vereis bepaalde datastrukture in die leksikografiese databasis en komplekse koppelvlakontwerp, beide items wat deur 'n multidissiplinêre span in die ontwerp van die woordeboek aangespreek moet word. Dit vorm dus deel van 'n struktuur wat gerig is op die maklike en suksesvolle onttrekking van inligting aan die data wat aan die beperkte artikels toegewys is. Vergelyk in hierdie verband figuur 3, die beperkte artikel wat vertoon word wanneer 'n gebruiker op die datamerker *Sinnverwandte Wörter* (= "betekenisverwante woorde") in figuur 2 klik:



Auge

Lesart: **'Sehorgan'**

Sinnverwandte Wörter

Beziehung(en) der Bedeutungsgleichheit/-äquivalenz

Synonym(e):

Sehorgan

Beziehung(en) des Bedeutungsgegensatzes

inkompatible(r) Partner:

Gesicht
Haar
Mund
Nase
Ohr

Die Auswahl der hier angegebenen Mitspieler richtet sich nach der statistischen Signifikanz im elexiko-Korpus und deckt diejenigen Körperteile ab, die in räumlicher Nähe zum *Auge* liegen.

Figuur 3: Uit *elexiko*

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1941 (Article)

Die data wat hier verstrek word, geld nie die ander betekenisonderskeidinge van die woord *Auge* nie. Die leksikograaf gebruik die datamerkers as subkommentaaruitbreidingsaanwysers (Wiegand en Gouws 2013: 296) wat die gebruiker help om by uitbreidings van die subkommentaar uit te kom. Ook hier kan die gebruiker geordende stappe volg ter onttrekking van die inligting — dit is deel van die inligtingsonttrekkingstruktuur van die betrokke woordeboekartikel.

Uit figuur 1–3 blyk dat 'n kommentaar — hier 'n semantiese kommentaar — 'n afwaarts uitgebreide artikelkonstituent is. In hierdie semantiese kommentaar kan daar afwaarts beweeg word na beperkte artikels wat verskillende vlakke beset. Binne 'n vlak, dit wil sê 'n enkele semantiese subkommentaar, is 'n horisontale beweging moontlik om by die data in die verskillende vakke uit te kom. Die horisontale beweging binne 'n vlak skep 'n soektonnel. 'n Soektonnel (Gouws 2022) is 'n struktuur met 'n geordende reeks stappe wat in 'n artikelvlak gevolg word om 'n aanduider in 'n bepaalde vak te bereik met spesifieke data gerig op die betrokke subkommentaar. Dit is byvoorbeeld die struktuur wat dit vir die gebruiker moontlik maak om vanaf die aanbieding in figuur 2 na die aanduider van 'n sinoniem in die vak vir betekenisverwante woorde in figuur 3 te beweeg.

Verskillende vlakke en vakke dwing die gebruiker (indien addisionele inligting benodig word) om verby die aanduiders wat met 'n eerste oogopslag sigbaar is in 'n artikel verder te beweeg na spesifieke kommentare en subkommentare. So 'n koppelvlakstruktuur kan ook inligtingsoorlading voorkom, deur die gebruiker slegs bloot te stel aan die data wat hulle benodig om hulle inligtingsbehoefte te bevredig, en nie aan die volledige stel komplekse data in die artikel nie. In hierdie proses begeef gebruikers hulle op soekprosedures met nuwe strukture soos uitgebreide kommentare en nuwe soekposisies soos soektonnels. Dit is middele wat die leksikograaf in die aanwending van die dataverspreidingstruktuur in werking stel. So 'n stelsel vereis wel dat die leksikograaf die gebruiker in aanmerking moet neem en moet verseker dat die nodige struktuur beskikbaar is om die soek, vind en onttrekking van inligting te verseker. Gebruikers moet bewus wees van die volle omvang van die data wat in 'n leksikografiese produk aangebied word en hulle moet daartoe in staat wees om die tersaaklike inligting te kan onttrek. Daarom moet leksikograwe nie net oor die dataverspreidingstruktuur besin nie, maar ook oor maniere om gebruikers daartoe in staat te stel om 'n gepaste woordeboekraadpleging en inligtingsonttrekking te kan deurvoer, in samewerking met die multidissiplinêre ontwikkelingspan van die woordeboek.

5. 'n Inligtingsonttrekkingstruktuur

Om suksesvolle woordeboekgebruik te verseker, is die aanbieding van die verlangde data nie genoeg nie. Strukture moet geskep word wat kan verseker dat die betrokke leksikografiese funksie bevredig word en wat die gebruiker kan

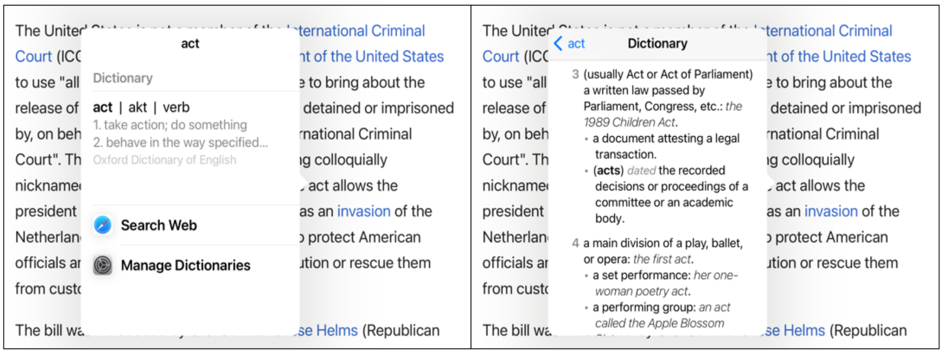
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1941 (Article)

help om die relevante data te vind, te interpreteer en op die regte manier te benut.

Om dit moontlik te maak, is die implementering van 'n inligtingsonttrekkingstruktuur noodsaaklik. So 'n struktuur bied 'n geordende reeks stappe wat dit moontlik maak om die gepaste inligting te onttrek aan die data wat in 'n woordeboekartikel geplaas is. 'n Woordeboek se inligtingsonttrekkingstruktuur blyk dikwels die spieëlbeeld van daardie woordeboek se dataverspreidingstruktuur te wees.

'n Inligtingsonttrekkingstruktuur kan verskillende inligtingsonttrekkingsroetes bepaal. In 'n woordeboek soos *elexiko* kan die gebruiker 'n uitvoerige toegangsroete benut om by die verlangde data uit te kom en die inligting te onttrek. 'n Gebruiker wat goed vertroud is met *elexiko* kan 'n korter toegangsroete volg deur bepaalde stappe in die toegangstruktuur uit te laat om gouer by die data te kom en die tersaaklike inligting te onttrek. Daar moet wel duidelike datamerkers wees om die soekroete te kan identifiseer en 'n optimale inligtingsonttrekking moontlik te maak.

E-toestelle bevat dikwels geïntegreerde woordeboeke wat uitgebreide of minder uitgebreide prosedures benut om inligting te onttrek. Gouws en Tarp (ter perse) gee aandag aan twee resultate waar 'n gebruiker van 'n e-toestel inligting aan die geïntegreerde woordeboek probeer onttrek. In albei gevalle is die gebruiker op soek na die betekenis van die naamwoord *act* soos dit in 'n spesifieke teks gebruik word. In die eerste geval klik 'n gebruiker wat 'n teks op 'n e-leser lees op die woord *act*. Die opskietvenster (figuur 4) vertoon 'n bewerking van dié woord vanuit die databasis wat aan die geïntegreerde woordeboek gekoppel is. Die eerste klik bied die eerste twee betekenisonderskeidinge van die woord *act* as werkwoord. Om by die bewerking van *act* as 'n naamwoord uit te kom, is 'n verdere klik nodig. Dan word die hele artikel vertoon:



act | akt | verb
1. take action; do something
2. behave in the way specified...
Oxford Dictionary of English

3 (usually Act or Act of Parliament) a written law passed by Parliament, Congress, etc.: *the 1989 Children Act*.
- a document attesting a legal transaction.
- (**acts**) *dated* the recorded decisions or proceedings of a committee or an academic body.

4 a main division of a play, ballet, or opera: *the first act*.
- a set performance: *her one-woman poetry act*.
- a performing group: *an act called the Apple Blossom*

Figuur 4: Uit die teks op die e-leser en die geïntegreerde woordeboek

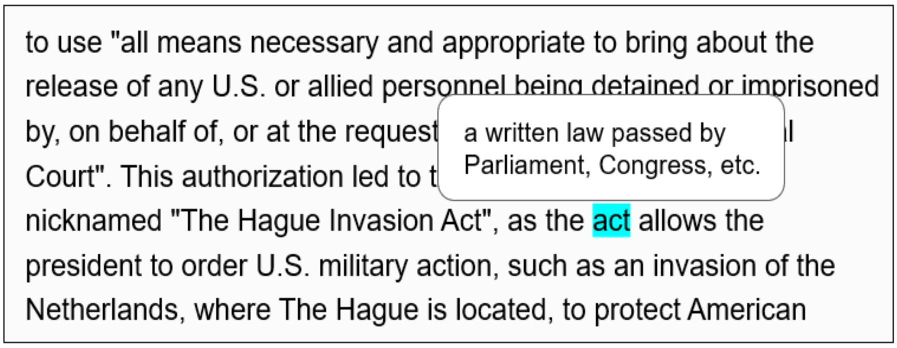
Die gebruiker kry steeds nie dadelik die regte betekenisparafrase nie en suksesvolle inligtingsonttrekking vereis dat die gebruiker na die tweede klik self 'n

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1941 (Article)

besluit moet maak oor watter betekenisparafrase die gepaste een is. Die inligtingsonttrekkingstruktuur is hier ondoeltreffend en lei nie tot 'n vinnige en ondubbelsinnige oplossing nie.

Vir verdere voorbeelde van problematiese soektogte op die Kindle-toepassing op 'n iPad, asook gewone lopende teks in 'n Google webblaaier, sien byvoorbeeld Bothma en Fourie (2024a en 2024b), Bothma en Gouws (2020 en 2022); vir voorbeelde van tipiese foute wat algemeen in die skakeling in die Kindle-toepassing voorkom, sien Bothma en Prinsloo (2013). Skakeling na die eerste lemma wat grafies ooreenkom met die woord waarop geklik word, het wye implikasies vir die gebruiker. Dit vereis dat die gebruiker noukeurig die volledige aanbod (wat moontlik oor meerdere artikels en verskeie skerms in die aanlyn woordeboek mag strek) moet deurlees om vas te stel wat die regte betekenis van die woord in konteks is. Dit kan heelwat tyd verspil. Dit impliseer verder dat die gebruiker 'n bepaalde minimum kennis van grammatika moet besit, om minstens die korrekte woordsoort in konteks te kan bepaal. In alle gevalle moet gebruikers die verskillende betekeninsopsies noukeurig evalueer om te besluit wat die korrekte betekenis in konteks is. Die ooraanbod van data — waarvan die meeste nie in die spesifieke geval relevant of korrek is nie — lei tot inligtingoorlading en frustrasie by die gebruiker (Gouws en Tarp 2017).

'n Regstreekse, verkorte en suksesvolle inligtingsonttrekkingsroete blyk uit 'n voorstel van Huang en Tarp (2021). 'n Relatief eenvoudige program laat byvoorbeeld skrywers, uitgewers, onderwysers, ensovoorts toe om op daardie woorde in 'n digitale teks te klik wat hulle as problematies vir die aangewese teikenlesers beskou. Die lemma word dan in die databasis en die geïntegreerde woordeboek nagegaan en slegs die gepaste betekenisonderskeiding en meegaande betekenisparafrase word aangedui. Wanneer die teikenleser dan op die betrokke woord in die teks klik, word hulle regstreeks na die verlangde betekenisparafrase gelei en 'n kitsonttrekking van die korrekte inligting is moontlik. Ter illustrasie: 'n klik op die naamwoord *act* in die bepaalde teks lei regstreeks tot die verklaring wat in figuur 5 se opskietvenster vertoon word:



to use "all means necessary and appropriate to bring about the release of any U.S. or allied personnel being detained or imprisoned by, on behalf of, or at the request [...] Court". This authorization led to t[...] nicknamed "The Hague Invasion Act", as the act allows the president to order U.S. military action, such as an invasion of the Netherlands, where The Hague is located, to protect American

a written law passed by Parliament, Congress, etc.

Figuur 5: Uit Huang en Tarp (2021)

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1941 (Article)

Dit is vernuwende en gebruikersgerigte leksikografie. 'n Bykomende merker, bv. "SEE MORE" kan ook in die opskietvenster geplaas word wat dit vir lesers met 'n behoefte aan bykomende inligting moontlik maak om nog inligting aan die databasis te onttrek. 'n Opskietspyskaart wat die beskikbare datatipes vertoon, gee die gebruiker die geleentheid om óf 'n enkele aanduider te kies óf te vra dat die volle artikel vertoon moet word. Hierdie benadering bied verskillende inligtingsonttrekkingsmoontlikhede omdat die inligtingsonttrekkingstruktuur verskillende onttrekkingsroetes kan bepaal. Soos aangetoon word in die hieropvolgende afdeling kan hierdie stelsel egter nie generies vir alle tekste beskikbaar wees nie, aangesien moontlike problematiese woorde in 'n bepaalde teks deur die skrywer/uitgewer/leksikograaf geïdentifiseer en geannoteer moet word, en leiding is net vir hierdie woorde beskikbaar. 'n Stelsel waarin die beginsel wat hier ter sprake is, naamlik direkte, outomatiese skakeling na die korrekte betekenis vir alle woorde in die teks en in die woordeboek, is wat vereis word.

Waar 'n klik op 'n woord in 'n teks die gebruiker na die volle artikel lei, volg die inligtingsonttrekkingsroete al die stappe van die volle inligtingsonttrekkingstruktuur. Waar 'n klik na 'n beperkte artikel lei, byvoorbeeld 'n semantiese subkommentaar of dalk slegs 'n glos as betekenisparafrase soos in figuur 5, is daar 'n verkorte inligtingsonttrekkingsroete wat net 'n gedeelte van die inligtingsonttrekkingstruktuur se stappe benut. Die behoeftes en vaardighede van die teikengebruiker is hier van bepalende belang en die leksikograaf moet vir meerdere moontlikhede voorsiening maak.

In figuur 1–5 was die fokus slegs op semantiese inligting en spesifiek die betekenisparafrase. Die onderliggende leksikografiese funksie wat hier bevredig moes word, is dié van teksbegrip. 'n Inligtingsonttrekkingstruktuur moet ook vir ander datatipes en ander funksies voorsiening kan maak. Dit is belangrik dat die leksikograaf die regte data aanbied en dat die gebruiker weet wat beskikbaar is en hoe en waar om dit te vind. 'n Gebruiker wat weet watter inligting onttrek kan word, sal weet hoe om regstreeks by slegs die betrokke aanduider uit te kom.

6. Bepaalde probleme

Die laasgenoemde opsie om 'n kitsonttrekking moontlik te maak, is nie sonder probleme nie en die leksikograaf moet hier met die nodige omsigtigheid te werk gaan. Suksesvolle begrip voorvereis dat die uitgewer, skrywer of onderwyser die teks moet annoteer. As nuwe tekste op dieselfde platform geplaas en toegang tot dieselfde databasis en geïntegreerde woordeboek bied, moet sulke tekste ook geannoteer word. Dié annotasie word gedoen op grond van dit wat die uitgewer, skrywer, onderwyser as problematies vir die teikenleser beskou. Dit is nie in die eerste plek die gebruiker se eie perspektief wat 'n rol speel nie en die beperkte bestek van die annotasies en die daargestelde inligtingsonttrekkingstruktuur gaan nie noodwendig die gebruiker help in al hulle probleemgevalle nie.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1941 (Article)

'n Alternatiewe werkswyse wat tot 'n inligtingsonttrekkingstruktuur kan lei wat alle woorde in 'n teks ter wille is, is waar die skakeling tussen die teks en die databasis konteksbewus moet wees. Die skakeling moet byvoorbeeld daartoe in staat wees om op grond van die woord waarop geklik word die regte woordsoortwaarde te kan identifiseer — in die geval van die bogenoemde optrede van *act* dat dit hier 'n naamwoord is en nie 'n werkwoord nie. Ter wille van 'n ondubbelsinnige inligtingsonttrekking moet die gepaste betekenisonderskeiding binne die konteks en koteks van die woord in die bepaalde teks geïdentifiseer kan word. In hierdie verband bied Inligtingkunde hulp aan die aanlyn leksikografie om met gesofistikeerde soek- en filtreertegnologie die regte afparing tussen 'n geklikte woord en die tersaaklike betekenisonderskeiding te verseker. Vergelyk in hierdie verband Bothma (2011), asook Bothma en Gouws (2022), Bothma en Fourie (2024a, b) en Tarp en Gouws (2020).

7. 'n Mensgerigte inligtingsonttrekkingstruktuur

Suksesvolle woordeboekgebruik is gewoonlik afhanklik van die naslaanvaardighede van die woordeboek se aangewese teikengebruikers. 'n Wesenlike deel van 'n tradisionele woordeboekkultuur, vergelyk Hausmann (1989), is dat die gemeenskap by die leksikografie moet aanpas en onder meer woordeboekvaardighede bemeester. Soos Bothma (2011) toon Tarp en Gouws (2020) ook aan dat die aanlyn omgewing nuwe moontlikhede geskep het om gebruikers op beter maniere by te staan. 'n Nuwe woordeboekkultuur wat onder andere op mensgerigte ontwerpbeginsels steun, kan die gebruik van naslaanbronne, insluitende woordeboeke en ander leksikografiese produkte, makliker maak vir gebruikers. Volgens Tarp en Gouws (2020: 2) plaas dit die verantwoordelikheid op die skouers van die leksikograaf om produkte te ontwerp wat intuïtiewe gebruik moontlik maak.

Die strewe na 'n verwerkliking van intuïtiewe gebruik geld ook ten opsigte van die ontwerp en implementering van 'n inligtingsonttrekkingstruktuur. Waar 'n mensgerigte inligtingsonttrekkingstruktuur gebruik word, lei 'n klik op 'n bepaalde woord in 'n teks die gebruiker onmiddellik na die aanduider wat die gebruiker as verstektipe geïdentifiseer het. Data vanuit die geïntegreerde woordeboek wat van die leksikografiese databasis afgelaai is, laat gebruikers toe om oplossings vir hulle probleme te vind sodat hulle feitlik sonder onderbreking verder kan lees. Die soeksone waarheen gebruikers gelei word, bevat geen aanduiders wat hulle nie nodig het nie en fokus, waar dit die verstekwaarde is, slegs op die betekenisparafrase ter bevrediging van 'n teksbegripfunksie. So 'n woordeboek beskik oor 'n kitsinligtingsonttrekkingstruktuur.

8. Kontekstualisering en skakeling

Intuïtiewe gebruik van 'n mensgerigte produk wat kitstoegang tot die aangebode data en 'n regstreekse inigtingsonttrekking bied, vereis slim e-toestelle en

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1941 (Article)

slim sagteware. Bothma en Gouws (2022) het reeds daarop gewys dat meer aandag gegee moet word aan skakelingprosedures tussen byvoorbeeld e-lesers, die onderliggende databasisse en die geïntegreerde woordeboeke. Die rol van die databasis raak in hierdie verband toenemend belangrik. Indien die gebruiker slegs 'n teksbegripbehoefte het, moet 'n klik op 'n woord in 'n teks op 'n e-leser nie 'n skakeling met die gepaste woordeboekartikel bewerkstellig nie, maar met die gepaste aanduider in die onderliggende databasis. Om dit moontlik te maak, is gesofistikeerde prosedures van kontekstualisering nodig.

Die skakelingsprosedure tussen 'n woord wat in 'n bepaalde konteks optree en die tersaaklike aanduider in die databasis voorveronderstel dinamiese artikels en leksikografiese strukture. Bothma en Gouws (2022) dui ten regte aan dat die geïntegreerde woordeboek moet aanpas by die buite-leksikografiese behoefte wat tot die betrokke naslaanhandeling gelei het; sien ook Bothma en Fourie (2024a, b). Suksesvolle skakeling vereis hoër kontekstualiseringsvereistes. Gekontekstualiseerde skakeling skep dan 'n verhouding tussen woorde in 'n e-teks en gebruiker-gespesifiseerde leksikografiese bronne. Die gebruiker word gelei na die presiese en relevante aanduiders en 'n gepersonifieerde inligtingsonttrekkingstruktuur word gevestig.

Suksesvolle skakeling verg die benutting van hoëvlaktegnologie. Slim sagteware, byvoorbeeld woordsoortmerkers, en die gebruik van menslike taaltegnologie, kunsmatige intelligensie en masjienleer moet in werking gestel word. Sodoende kan 'n gegewe teks op die e-leser geïnterpreteer word en na wisselwerking met die databasis kan die gebruiker van 'n gekontekstualiseerde resultaat voorsien word. Die toepassing van mensgerigte ontwerpbeginsels bring mee dat die sagteware geen, of hoogstens beperkte, manipulering deur die gebruiker nodig het. Die inligtingsonttrekkingstruktuur vertoon homself as 't ware aan die gebruiker.

Kunsmatige intelligensie — beide "tradisionele" kunsmatige intelligensie, wat tipies van natuurliketaalprosessering (NLP, "natural language processing") gebruik sou kon maak, en generatiewe kunsmatige intelligensie, gebaseer op groottaalmodelle (LLMs, "large language models") — kan heel moontlik 'n belangrike bydraende rol speel in die kontekstualiseringsproses. Dit word egter nie in hierdie artikel bespreek nie, omdat die fokus nie hier op die implementering is van die voorstelle wat hier gemaak word nie. Verdere navorsing sou met vrug hieraan kon aandag gee.

9. Ten slotte

Leksikograwe moet vernuwend dink en beplan, en poog om die gebruiker die geleentheid te bied om moeiteloos 'n optimale onttrekking van inligting te kan bewerkstellig. In die beplanning van 'n inligtingsonttrekkingstruktuur moet leksikograwe nie die artikel as mikpunt van 'n soektog sien nie, maar eerder 'n spesifieke aanduider wat daardie data bied waarna 'n spesifieke gebruiker in 'n spesifieke gebruiksituasie en met inagneming van 'n spesifieke konteks op soek is.

Data en maklike toegang tot data is die goudstandaard van die moderne leksikografie en daarom mag die rol van die databasis waaraan 'n geïntegreerde woordeboek gekoppel is, nooit onderskat word nie. Trouens, dit plaas die beplanning van die leksikografiese databasis op die voorgrond, sien in hierdie verband ook Fuertes-Olivera et al. (2018) en Tarp (2022). Die leksikograaf moet besin oor beide die dataverspreidingstruktuur en die inligtingsonttrekkingstruktuur, en dan bepaal hoe die leksikografiese databasis gekonstrueer moet word om optimaal aan gebruikers se inligtingsbehoeftes te kan voldoen. So 'n leksikografiese databasis vereis 'n hoë partikelaard om te verseker dat data maklik in vlakke en vakke gegroepeer kan word. Die ontwerp en implementering van so 'n komplekse databasis vereis uiteraard die insette van die leksikograaf (soos reeds aangedui), maar aangesien die leksikograaf in die meeste gevalle nie ook 'n programmeerder of stelselontwikkelaar is nie, word die insette van rekenaarwetenskaplikes vereis. Die belangrikheid van 'n mensgerigte koppelvlak is ook reeds aangetoon. Die leksikograaf is uiteraard ook nie 'n koppelvlakontwerper nie, en moet dus met spesialiste op die gebied van mens–rekenaarinteraksie (HCI = "human computer interaction") en gebruikerservaring (UX = "user experience") saamwerk om die koppelvlak te ontwerp. Dit beklemtoon die noodsaaklikheid van 'n multidissiplinêre span om die leksikografiese databasis en die beplande dataverspreidingstrukture en inligtingsonttrekkingstrukture te ontwerp. Indien hierdie beplanning nie vooraf in diepte gedoen word nie, sal dit onmoontlik wees om in 'n latere stadium gesofistikeerde inligtingsonttrekkingstrukture te implementeer wat aan 'n verskeidenheid gebruikersbehoeftes voldoen.

In die leksikografie was daar 'n verskuiwing vanaf diskrete teks na taaldata en 'n beweging van 'n gerigtheid op die publikasie van boeke na 'n voorsiening van taaldienste (Ogilvie 2021: 91). Hierdie taaldienste sluit die erkenning van verskillende leksikografiese funksies asook die daarstelling van 'n verskeidenheid datatipes in. Dit is gebaseer op 'n mensgerigte benadering wat intuïtiewe woordeboekgebruik moontlik maak. Daarom moet die taaldienste op so 'n manier voorsien word dat gebruikers op 'n ondubbelsinnige en regstreekse manier daardie data kan bereik waaraan hulle die benodigde inligting kan onttrek. Dit kan slegs gebeur as 'n woordeboek se dataverspreidingstruktuur aangevul word met 'n inligtingsonttrekkingstruktuur.

Bronnelys

Aanlyn woordeboeke

elexiko: https://www.owid.de/docs/elex/start.jsp

Gedrukte woordeboeke

Hartmann, R.R.K. en G. James. 1998. *Dictionary of Lexicography.* Londen/New York: Routledge.
Johnson, S. 1755. *Dictionary of the English Language*. London: J. & P. Knapton.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1941 (Article)

Ander bronne

Bergenholtz, H., S. Tarp en H.E. Wiegand. 1999. Datendistributionsstrukturen, Makro- und Mikrostrukturen in neueren Fachwörterbüchern. Hoffmann, L. et al. (Reds.). 1999. *Fachsprachen. Ein internationales Handbuch zur Fachsprachenforschung und Terminologiewissenschaft/Languages for Special Purposes. An International Handbook of Special-Language and Terminology Research, Bd./Vol.* 2: 1762-1832. Berlyn: De Gruyter.

Bothma, T.J.D. 2011. Filtering and Adapting Data and Information in the Online Environment in Response to User Needs. Fuertes-Olivera, P.A. en H. Bergenholtz (Reds.). 2011. *e-Lexicography: The Internet, Digital Initiatives and Lexicography*. Londen/New York: Continuum: 71-102.

Bothma, T.J.D. en I. Fourie. 2024a. Contextualised Dictionary Literacy, Information Literacy, and Information Behaviour in the e-Environment. *Library Management* (ter perse).

Bothma, T.J.D. en I. Fourie. 2024b. 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.

Bothma, T.J.D. en R.H. Gouws. 2020. e-Dictionaries in a Network of Information Tools in the e-Environment. *Lexikos* 30: 29-56.

Bothma, T.J.D. en 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.

Bothma, T.J.D. en D.J. Prinsloo. 2013. Automated Dictionary Consultation for Text Reception: A Critical Evaluation of Lexicographic Guidance in Linked Kindle e-Dictionaries. *Lexicographica* 29: 165-198.

Fuertes-Olivera, P.A., S. Tarp en P. Sepstrup. 2018. New Insights in the Design and Compilation of Digital Bilingual Lexicographical Products: The Case of the Diccionarios Valladolid-UVa. *Lexikos* 28: 152-176.

Gouws, R.H. 2014. Article Structures: Moving from Printed to e-Dictionaries. *Lexikos* 24: 155-177.

Gouws, R.H. 2022. Extended Article Comments in Online Dictionaries. *Lexicography. Journal of ASIALEX* 9(2): 111-138.

Gouws, R.H. 2023. Nuwe raamstruktuurtipes in aanlyn woordeboeke. *Lexikos* 33(1): 382-403.

Gouws, R.H. et al. (Reds.). 2013. *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography.* Berlyn: De Gruyter.

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. Ter perse. Despite Current Challenges: Lexicography Has a Bright Future. *Lexicographica* 40.

Hartmann, R.R.K. 1989. Sociology of the Dictionary User: Hypotheses and Empirical Studies. Hausmann, F.J. et al. (Reds.). 1989–1991:103-111.

Hausmann, F.J. 1989. Die gesellschaftlichen Aufgaben der Lexikographie in Geschichte und Gegenwart. Hausmann, F.J. et al. (Reds.). 1989–1991: 1-19.

Hausmann, F.J. et al. (Reds.). 1989–1991. *Wörterbücher. Dictionaries. Dictionnaires. An International Encyclopedia of Lexicography.* Berlyn: De Gruyter.

Huang, F. en S. Tarp. 2021. Dictionaries Integrated into English Learning Apps: Critical Comments and Suggestions for Improvements. *Lexikos* 31(1): 68-92.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1941 (Article)

Ogilvie, S. 2021. The Future of Dictionaries and Lexicography. *Dictionaries: Journal of the Dictionary Society of North America* 42(2): 91-94.

Tarp, S. 2008. *Lexicography in the Borderland between Knowledge and Non-knowledge: General Lexicographical Theory with Particular Focus on Learner's Lexicography*. Berlyn/New York: Max Niemeyer.

Tarp, S. 2022. Turning Bilingual Lexicography Upside Down: Improving Quality and Productivity with New Methods and Technology. *Lexikos* 32(1): 66-87.

Tarp, S. en R.H. Gouws. 2020. Reference Skills or Human-Centered Design: Towards a New Lexicographical Culture. *Lexikos* 30: 470-498.

Wiegand, H.E. 1989. Arten von Mikrostrukturen im allgemeinen einsprachigen Wörterbuch. Hausmann, F.J. et al. (Reds.). 1989–1991: 462-501.

Wiegand, Herbert Ernst. 1998. *Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie.* Berlyn/New York: De Gruyter.

Wiegand, H.E., S. Beer en R.H. Gouws. 2013. Textual Structures in Printed Dictionaries. An Overview. Gouws, R.H. et al. (Reds.). 2013: 31-73.

Wiegand, H.E. en R.H. Gouws. 2013. Addressing and Addressing Structures in Printed Dictionaries. Gouws, R.H. et al. (Reds.). 2013: 273-314.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1942 (Article)

L'Evolution de la Terminologie de la Plasturgie entre 1963–2018: Analyse Diachronique et Synchronique

Valentina-Nicoleta Văsioiu, *Lucian Blaga University of Sibiu, Romania (valentinanicoleta.vasioiu@ulbsibiu.ro)*
(https://orcid.org/0009-0007-3273-6636)
et
Marilena Milcu, *Lucian Blaga University of Sibiu, Romania (maria.milcu@ulbsibiu.ro)*
(https://orcid.org/0000-0002-4221-8855)

Résumé: Le présent article se concentre sur la description du langage technique, en particulier la terminologie de la plasturgie du point de vue diachronique et synchronique, par l'intermédiaire des analyses comparatives réalisées à partir du lexique usité dans les années 1960 et 2010. Nous avons recours aux investigations effectuées à l'aide du logiciel *QI Macros,* des investigations qui nous ont permis d'obtenir des mesures statistiques pour les structures morphologiques identifiées (*syntagmes nominaux SN, verbaux SV* et *adjectivaux SA*) et un classement de termes candidats selon leur productivité (*les TC caractérisés par une forte productivité N_S63 < N_S18, les TC caractérisés par symétrie N_S63 = N_S18* et *les TC caractérisés par involution N_S63 > N_S18).* Nous nous intéressons aussi à la représentation du lexique transdisciplinaire (Jacquey et al. 2018) ainsi qu'au vocabulaire de la plasturgie, notamment les vocabulaires spécifiques pour la physique, la chimie, la gestion de la production et la mécanique.

Mots clés: LEXIQUE TECHNIQUE, TERMINOLOGIE DE LA PLASTURGIE, UNITÉ TERMINOLOGIQUE, SYNTAGME, PRODUCTIVITÉ, TERME CANDIDAT *TC*

Abstract: The Evolution of the Plastics Terminology between 1963 and 2018: A Diachronic and Synchronic Analysis. Our article focuses on the technical language description, particularly, it concentrates on the study of plastic terminology, from a diachronic and synchronic point of view, through the intermediary of a comparative analysis starting from the plastic vocabulary used in the 1960s and 2010s. We have conducted linguistic investigations with the help of the software *QI Macros,* investigations which helped us obtain statistical measurements for the identified morphological structures (*nominal phrases NP, verbal phrases VP* and *adjectival phrases AP*) and a classification for the candidate terms according to productivity criteria (*TC with a high productivity N_S63 < N_S18, TC characterized by its symmetry N_S63 = N_S18, as well as TC characterized by its involution N_S63 > N_S18).* We have also examined the representation of transdisciplinary vocabulary[1] (Jacquey et al. 2018) within the plastic terminology, especially vocabularies specific for fields such as physics, chemistry, production management and mechanics.

Lexikos 34 (AFRILEX-reeks/series 34: 2024): 371-396

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1942 (Article)

Keywords: TECHNICAL VOCABULARY, PLASTIC TERMINOLOGY, TERMINOLOGICAL UNIT, SYNTAGM, PRODUCTIVITY, TERM CANDIDATE

1. Introduction

La terminologie technique a toujours suscité beaucoup d'intérêts parmi les linguistes et à l'époque actuelle aussi (Zanola 2021; Forner et Thörle 2016; Avornicesei et al. 2021; Bidu-Vrânceanu 2000, 2007; Ilinca et Tomescu 2013; Tomescu 2022 etc.), puisqu'elle se trouve «au service de la divulgation de la science, de la technique et des activités spécialisées» (Cabré 2016: 74). Dans cet article, nous avons envisagé d'explorer, comme nombre de théoriciens avant nous, la terminologie technique (Laroche et al. 2011; Arhire 2014; Zufferey 2020), notamment le lexique de l'industrie du plastique en réalisant un corpus d'analyse à partir du vocabulaire usité dans les années 1963–2018 et des analyses comparatives qui nous ont offert une vue d'ensemble, globale, sur l'évolution ou l'involution des unités identifiées. L'enrichissement de la terminologie de l'industrie du plastique est inéluctable et il est «dû à la nécessité de la nomination de nouveaux phénomènes» (Halyan 2014: 38) physiques et chimiques, de nouveaux outillages et robots, des procédés de fabrication et des propriétés de matériaux plastiques. Les innovations apparues et leur applicabilité entrainent d'un côté des changements au niveau morphologique (Janssen et Van Campenhoudt 2005) et de l'autre côté déterminent l'élimination de certaines unités terminologiques et la croissance ou la réduction du degré d'utilisation des autres unités au sein des milieux professionnels.

La première partie de l'article présente les travaux sur le lexique de la plasturgie afin de mettre en valeur la similitude ou la complémentarité de notre étude. Plus précisément, elle nous révèle les travaux qui renvoient au thème spécifique de la terminologie du plastique et des polymères. La deuxième partie de l'article introduit la méthodologie quantitative (N'Da 2015) et les instruments de travail utilisés afin de repérer les structures spécifiques et d'établir, à partir des mesures statistiques générées par le logiciel *QI Macros*, des critères de distinction concernant la productivité de termes candidats.

2. Objectifs et hypothèses

2.1 Objectifs

2.1.1 Dans cet article, nous analysons les structures morphologiques de type *SN, SV* et *SA* du point de vue linguistique (Pierre et Parth 2005; Polzin-Haumann et Schweickard 2015; Zufferey et Moescheler 2015). Les structures ont été identifiées dans les deux vocabulaires considérés comme points de repère, à savoir Lexicon 1963 et ISO 2018. C'est à travers les analyses comparatives que nous obtenons des informations essentielles pour «mieux comprendre les phénomènes relatifs à la terminologie actuellement en usage» (Dankova 2021: 26).

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1942 (Article)

2.1.2 L'étude du «lexique transdisciplinaire dans le processus de la terminologisation» (Jacquey et al. 2018: 27) nous intéresse dans la mesure où les lexiques de la physique, de la chimie et de la mécanique montrent une présence importante au sein du vocabulaire de la plasturgie. Toutefois, on se propose d'examiner l'influence indéniable des termes appartenant au lexique transdisciplinaire dans la constitution des nouvelles structures.

2.2 Hypothèses

2.2.1 La première hypothèse est qu'il n'y a pas de linéarité en ce qui concerne le lexique technique et que l'élimination de certaines unités terminologiques constitue un phénomène naturel, comme la croissance ou la réduction du degré d'utilisation des autres unités au sein des milieux professionnels.

2.2.2 La deuxième hypothèse est que certaines unités terminologiques préexistantes montrent une plus grande tendance à l'évolution en vue de répondre aux besoins communicationnels. Il est à noter que les techniques utilisées dans les entreprises spécialisées dans la production des produits en plastique sont soumises aux modifications constantes entraînées par le progrès technique et scientifique et bien sûr par l'adaptation de la production à la demande du marché, tout comme le souligne l'Association Plastics Europe: «Les plastiques sont à l'origine d'un éventail considérable et en expansion constante d'innovations qui contribuent au développement durable, à la sécurité, à l'allongement de l'espérance de vie et à de meilleures performances» (Plastics Europe [legacy.plasticseurope], n.d.).

2.3 Analyse de la littérature de spécialité

Afin de mettre en relief les dernières études menées par les chercheurs dans le champ de la terminologie du domaine spécialisé étudié, nous avons examiné les travaux sur la terminologie du plastique, en utilisant plusieurs plateformes et moteurs de recherche: le moteur de recherche *Google Academic*, la plateforme *CEEOL (Central EasternEuropean Online Library)* et *ERIHPLUS (European Reference Index for the Humanities and Social Sciences)*.

2.3.1 *Le moteur de recherche Google Academic* a été utilisé pour l'étude du sujet *analyses / études de la terminologie de la plasturgie.* Pour y parvenir, nous avons décidé d'introduire dans le moteur de recherche quatre structures différentes pour identifier les articles et les livres parus sur le sujet qui nous intéresse. Nous restreignons la période de recherche à 2015–2023. Cela devrait amener à retrouver les travaux les plus actuels dans le domaine de la plasturgie et de la linguistique, en particulier la terminologie. Les structures introduites sont les suivantes: *terminologie de la plasturgie, vocabulaire de la plasturgie, analyse de la terminologie des polymères, étude linguistique de la terminologie des polymères. La première recherche* a été réalisée par l'introduction de la structure *terminologie de la plasturgie,* en sélectionnant la période 2015–2023 dans le moteur de recherche Google Academic et nous avons obtenu **142** résultats: 0 articles sur la terminologie de la plasturgie, 5 articles pertinents, 46 articles sur l'industrie de la plasturgie et 91 articles sans rapport évident avec notre sujet. *La deuxième recherche* a été réalisée par l'introduction de la structure *vocabulaire de la plasturgie,* en sélectionnant la période 2015–2023 dans le moteur de recherche Google Academic et nous avons obtenu **172** résultats: 0 articles pertinents, 34 articles sur l'industrie de la plasturgie et 138 articles sans rapport évident avec notre sujet. *La troisième recherche* a été réalisée par l'introduction de la structure *analyse de la terminologie des polymères,* en sélectionnant la période 2015–2023 dans le moteur de recherche Google Academic et nous avons obtenu **3650** résultats: 5 articles pertinents, articles sur les polymères et articles sans rapport avec notre sujet. *La quatrième recherche* a été réalisée par l'introduction de la structure *étude linguistique de la terminologie des polymères,* en sélectionnant la période 2015–2023 dans le moteur de recherche Google Academic et nous avons obtenu: 4 articles pertinents, 29 articles sur les polymères, 79 articles sur terminologie / linguistique et 35 sans rapport avec notre sujet.

Après la recherche sur Google Academic, nous avons retrouvé **4** articles pertinents après l'introduction de la structure *étude linguistique de la terminologie des polymères* et **5** articles pertinents après l'introduction de la structure *étude linguistique de la terminologie des polymères.* Afin de juger leur pertinence, les articles ont été analysés pour vérifier si leur contenu est adéquat pour notre sujet. Seulement 6 articles sont retenus comme appropriés. Dans le tableau suivant nous avons classé les articles selon le nombre de citations:

Tableau 1: Articles pertinents identifiés à travers la recherche sur Google Academic

	Titre	Auteur	Année	Nombre de citations
1.	*Terminologie des emballages et des matériaux commençant par le préfixe «bio»*	Marie Berteloot	2019	3
2.	*La néologie terminologique en français dans le domaine des fibres chimiques*	Klara Dankova	2021	0
3.	*La composition dans le lexique textile*	Angelica Preda	2019	0
4.	*Étude terminologique de la chimie en arabe dans une approche de fouille de textes.*	Albeiriss Baian	2018	0
5.	*De la terminologie textile: fibres et fils*	Silvia Pitiriciu	2016	0
6.	*Traduction de termes textiles français en suédois: — traduction d'un extrait du livre Textiles techniques et fonctionnels, matériaux du XXIe siècle*	Petra Ivarsson	2015	0

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1942 (Article)

Nous pouvons observer facilement que les articles de spécialité retrouvés se rapprochent plus des secteurs connexes au domaine de la plasturgie, parmi lesquels des recherches portant sur la terminologie des emballages, des fibres chimiques, de la chimie, la terminologie textile, mais aucune étude sur la terminologie du plastique n'a été identifiée avec le moteur de recherche Google Academic.

2.3.2 Nous avons poussé plus loin le processus d'analyse de la littérature de spécialité en initiant une recherche sur le site *Central and Eastern European Online Library* (https://www.ceeol.com). L'introduction des structures différentes a facilité l'identification des articles et des livres parus sur le sujet qui nous intéresse, en commençant par des structures qui renvoient au thème général de la terminologie et en continuant avec le thème plus spécifique de la terminologie du plastique ou des polymères. Dans la première recherche, nous avons introduit la structure *terminologie* et ensuite *Applied Linguistics* et nous avons obtenu **167** résultats. Malheureusement, il n'y a pas de filtre pour choisir la période. Un travail de tri est réalisé en analysant les résultats et on trouve des articles et des ouvrages portant sur la terminologie et la didactique, la terminologie et la traduction, le français de spécialité, la terminologie médicale, la terminologie juridique, la terminologie de la gastronomie, la terminologie de la pharmacologie, du marketing, de l'écologie, de la psychanalyse etc., mais aucun résultat pertinent sur la terminologie de la plasturgie. Dans la deuxième recherche, nous avons introduit la structure *terminologie,* en cochant la *langue française* et ensuite *Terminology* et nous avons obtenu **50** résultats. À nouveau, il n'y a pas de filtre pour choisir la période. Le travail de tri est réalisé en analysant les résultats et on trouve des articles et des ouvrages portant sur la terminologie technique et scientifique, les langues spécialisés, la linguistique, la traduction, la terminologie électronique, du sport, de la bioéthique, des arts etc., mais aucun résultat sur la terminologie de la plasturgie. Cependant, nous avons trouvé un article lié au domaine de la plasturgie, plus précisément un article qui renvoie au domaine de la physique:

Tableau 2: Articles pertinents identifiés à travers la recherche sur *Central and EasternEuropean Online Library*

	Titre	Auteur	Année	Nombre de citations
1.	*La structure et la sémantique des groupes de mots termes physiques de français*	Oksana Halyan	2014	0

Dans la troisième recherche, nous avons introduit la structure *terminologie plastique,* et ensuite *Applied Linguistics* et nous avons obtenu **8** résultats, mais aucun sur la terminologie de la plasturgie. Dans la quatrième recherche, nous avons introduit la structure *terminologie polymères,* en cochant la *langue française* et nous avons obtenu **3** résultats. Deux articles ont été sans rapport avec notre sujet, mais le troisième apporte des informations sur les fibres chimiques, notamment les polymères avec des applications dans le domaine textile:

Tableau 3: Articles pertinents identifiés à travers la recherche sur *Central and Eastern European Online Library*

	Titre	**Auteur**	**Année**	**Nombre de citations**
1.	*La néologie terminologique en français dans le domaine des fibres chimiques*	Klara Dankova	2021	0

2.3.3 *ERIH PLUS.* En vue d'initier une recherche sur cette plateforme, nous avons choisi la période *2015–2023* et la catégorie *Linguistics* et nous avons introduit la structure *terminologie plastique.* **21** résultats ont été obtenus, parmi lesquels des articles sur la terminologie du vin, du Covid 19, sur la sémantique etc., mais aucun résultat pertinent sur notre sujet. *Une deuxième recherche* a été menée en introduisant la structure *terminologie polymères*, en préservant la période *2015–2023* et la catégorie *Linguistics.* Nous avons obtenu **1** résultat qui porte sur l'enseignement des langues de spécialité en France et en Pologne, donc aucune relation avec notre sujet. Dans la troisième recherche, nous avons introduit la structure *vocabulaire plastique* et nous avons obtenu **28** résultats portant sur la sémantique, sur la traduction, la terminologie du vin etc., mais aucun résultat pertinent.

En conclusion, il est à noter que le domaine de la terminologie technique suscite l'intérêt des chercheurs et que des études sur la terminologie technique ont été rédigées, notamment des travaux menés dans des domaines comme l'électronique, le sport, la bioéthique, les arts, la physique, la chimie, le domaine textile et le domaine des emballages bio. Les sujets abordés incluent la néologie, la composition, la traduction de termes techniques, des analyses sur la sémantique des termes physiques.

2.4 Limites de la recherche

En premier lieu, le format papier du Lexicon 1963 a constitué un facteur décisif dès le début de notre étude, puisque c'était déjà évident que pour l'introduction manuelle des unités terminologiques recensées dans cet ouvrage et pour leur recherche dans le document de type word ISO 2018, nous aurons besoin d'une période étendue. En second lieu, l'erreur humaine concernant l'introduction manuelle des unités terminologiques peut apparaitre. De plus, nous sommes obligés de reconnaitre que le Lexicon 1963 n'est pas exhaustif et qu'il y a la possibilité que d'autres unités terminologiques usitées dans les années 1960 aient été exclues pour divers motifs.

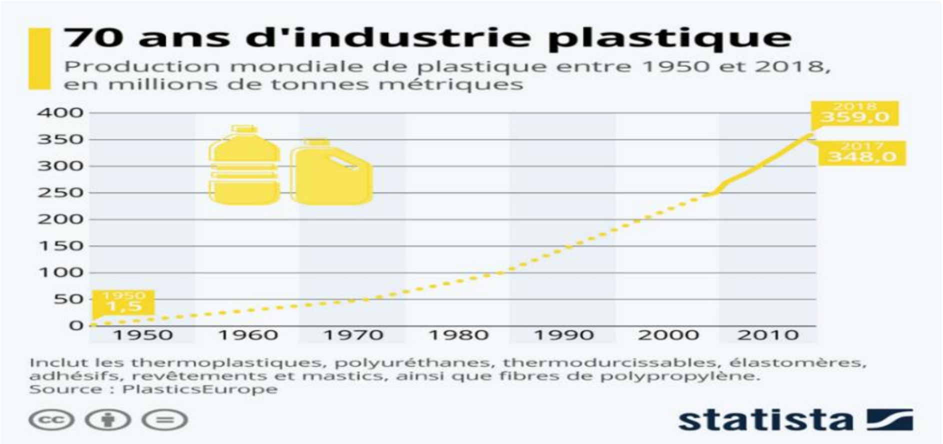
3. Méthodologie et instruments de travail

Afin de pouvoir vérifier nos hypothèses et atteindre nos objectifs, nous avons réalisé des travaux terminologiques focalisés sur le lexique de la plasturgie dans sa globalité. La méthodologie adoptée dans nos travaux est quantitative et nous avons travaillé dans une fenêtre temporelle longue, allant des années 1960 jusqu'à 2018. La collecte des données terminologiques a été réalisée à partir de ressources suivantes:

— L'ouvrage *Plastics Lexicon — Processing and Machinery in Six Languages*, Elsevier, 1963, ci-après dénommé **Lexicon 1963**;
— Le standard international *ISO 472:2013 — valable aussi pour 2018, (fr) Plastiques — Vocabulaire*, ci-après dénommé **ISO 2018**.

Le choix des ressources a été déterminé par le désir de réunir des données terminologiques représentatives pour le secteur de la plasturgie. La première ressource a été retrouvée pendant une recherche documentaire sur la terminologie du plastique et le livre inclut des termes «basic to the technology and related activities of the plastics industry» (Wittfoht 1963). La deuxième ressource a été choisie puisqu'elle fournit «la terminologie la plus actuelle utilisée dans les milieux professionnels» (Dankova 2021: 98). Nous avons choisi ces vocabulaires parce que les années 1960 constituent le début de l'évolution fulminante de l'industrie plastique et les années 2010 l'apogée, ainsi que nous le présente la figure 1. Par conséquent, les deux ouvrages peuvent être considérés comme des repères dans le progrès de la terminologie du plastique et leur exploitation nous ont offert l'opportunité de réaliser une analyse diachronique et synchronique.

Figure 1: L'essor de l'industrie plastique depuis les années 1950 (Gaudiaut 2020)



70 ans d'industrie plastique

Production mondiale de plastique entre 1950 et 2018, en millions de tonnes métriques

Inclut les thermoplastiques, polyuréthanes, thermodurcissables, élastomères, adhésifs, revêtements et mastics, ainsi que fibres de polypropylène.
Source : PlasticsEurope

3.1 La préparation des unités terminologiques

L'ouvrage Lexicon 1963 se retrouve en format papier et en vue de pouvoir travailler avec les termes proposés, nous avons introduit manuellement **1360 unités terminologiques,** en français, dans un fichier de type *.xls.*, ci-après dénommé **fichier d'analyse**, qui ont été rangées en ordre alphabétique. Le fichier d'analyse, qui sera soumis aux investigations avec un logiciel spécialement conçu pour le format Excel, contient 4 colonnes. Un étiquetage morphologique (Vidal-Gorène et al. 2020) a été effectué pour chaque unité dans la colonne 2 et nous avons introduit l'unité identifiée dans le vocabulaire de l'ISO 2018 dans la colonne 3 et le type de structure a été marqué dans la colonne 4. Dans le tableau suivant nous avons inséré un fragment du fichier d'analyse. Le tableau 4 nous propose des unités terminologiques indexées dans le Lexicon 1963 et les unités équivalentes identifiées dans l'ISO 2018. Pour chaque unité on a réalisé une annotation manuelle concernant la structure morphologique.

Tableau 4: Fragment du fichier d'analyse qui a été soumis à l'exploitation à l'aide du logiciel *QI Macros*

Lexicon 1963	**Structure morphologique S_63**	**2018 ISO- Vocabulaire plastique**	**Structure morphologique S_18**
agent antistatique	N+adj	*agent de pontage*	N+prep+N
agent d’accrochage	N+prep+N	*agent anti adhérent*	N+adj
agent de démoulage	N+prep+N	*agent de durcissement*	N+prep+N
agent de séparation	N+prep+N	*agent bloqué*	N+adj
		agent inhibé	N+adj
		agent d’expansion	N+prep+N
		agent chimique	N+adj
		agent de traitement	N+prep+N
		agent antifloculant	N+adj
		agent émulsionnant	N+adj
		agent poisseux	N+adj
		agent de démoulage	N+prep+N
		agent de réticulation chimique	N+prep+N+adj
adhésif en feuille	N+prep+N	*adhésifséché*	N+adj
		adhésifséparé	N+adj
		adhésif sensible	N+adj
		adhésifanaérobic	N+adj

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1942 (Article)

		adhésif à durcissement à froid	N+prep+N+prep+N
		adhésif à prise à température ambiante	N+prep+N+prep+N+ adj
		adhésif à prise à froid	N+prep+N+prep+adj
		adhésif conducteur	N+adj
		adhésif de contact	N+prep+N
		adhésif en émulsion	N+prep+N
		adhésif encapsulé	N+adj
		Adhésif en film	N+prep+N
		adhésif expansible in situ	N+adj+prep+N
		adhésif à joint épais	N+prep+N+adj
		adhésif thermocollant	N+adj
		adhésif thermofusible	N+adj
		adhésif à prise à chaud	N+prep+N+prep+adj
		adhésif durcissant par humidification	N+adj+prep+N
		adhésif multicomposant	N+adj
		adhésif simple face	N+adj+N
		adhésif en pâte	N+prep+N
		adhésif plastisol	N+adj
		adhésif sensible à la pression	N+adj+prep+N
		adhésif de réaction	N+prep+N
		Adhésif réactif	N+adj
		adhésif thermodurcissable	N+adj
		adhésif autopolymerisable	N+adj
		adhésif à application séparée	N+prep+N+adj
		adhésif réactivable par un solvant	N+adj+prep+N
		adhésif à base de solvant	N+prep+N+prep+N
		adhésifen solution	N+prep+N
		adhésif sans solvant	N+prep+N
		Adhésif en aérosol	N+prep+N
		adhésif structural	N+adj
		adhésif en film avec support	N+prep+N+prep+N
		Adhésif durci	N+adj

		adhésif bicomposant	N+adj
		adhésif biface	N+adj
		adhésif en film sans support	N+prep+N+prep+N
		adhésif à base d'eau	N+prep+N+prep+N
		adhésif aqueux	N+adj
		adhésif de rétention d'eau	N+prep+N+prep+N
angle de chamfrain	N+prep+N	*angle droit*	N+adj
angle de depouille	N+prep+N	*angle de phase*	N+prep+N
angle d'ouverture	N+prep+N	*angle d'oscillation*	N+prep+N
angle de pliaje	N+prep+N	*angle égale*	N+adj
		angle de depart	N+prep+N
		angle de chute	N+prep+N
		angle de contact	N+prep+N
		à angle vif	N+adj
		tête d'angle	N+prep+N
assemblage par vis	N+prep+N	*temps d'assemblage*	N+prep+N
		rupture d'assemblage	N+prep+N
baguette d'apport	N+prep+N	*baguette d'apport*	N+prep+N
bande continue de papier	N+adj+prep+N	*bande chauffante*	N+adj
barreau d'essai normalisé	N+prep+N+adj	*barreau de traction*	N+prep+N
bavure	N	*bavure*	N
Bloc	N	*bloc*	N
Bois comprimé	N+adj	*bois de cellulose au sulfate*	N+prep+N+prep+N

Le tableau 4 nous permet de générer des observations encore partielles: (a) certaines unités terminologiques indexées dans le Lexicon 1963 ont été retrouvées sous la même forme dans l'ISO 2018 (*agent de démoulage, bavure, bloc, baguette d'apport etc.*) ou sous formes différentes (*agent → agent de pontage, agent anti adhérent; assemblage par vis → temps d'assemblage*, *rupture d'assemblage* etc.) et (b) d'autres unités n'ont pas été retrouvées (*agent d'accrochage, agent de séparation* etc.).

3.2 Les instruments de travail

Notre analyse se réalise en utilisant les données introduites manuellement dans le fichier d'analyse de type Excel, qui contient deux listes: la liste des unités terminologiques recensées dans le Lexicon 1963 et celle de leurs équivalents retrouvés dans l'ISO 2018. Par conséquent, nous avons besoin d'un logiciel spécialement conçu pour Excel afin de générer des graphiques plus complexes. Après une recherche sur les logiciels les plus appropriés, *QI Macros SPC SOFTWARE FOR EXCEL, Any Count, Analyse-it, Analysis Tool Pak* ont été identifiés. Le logiciel *QI Macros SPC SOFTWARE FOR EXCEL* a retenu notre attention car il peut être utilisé pour créer des tableaux et des graphiques complexes à partir des informations introduites dans un fichier format .xls et il offre une multitude de représentations graphiques et des diagrammes. Le logiciel est disponible sur le site *https://www.qimacros.com* et de plus, il peut être utilisé directement dans notre fichier .xls en cliquant sur *QI Macros*, situé après *View.* Le logiciel a été appliqué de façon systématique et il nous a permis de réaliser des investigations diverses qui seront présentées et discutées dans la section suivante.

4. Analyse lexicale des unités terminologiques représentatives du domaine de la plasturgie

En vue d'opérer des analyses lexicales sur les syntagmes indexés (syntagmes nominaux, verbaux, adjectivaux) dans le Lexicon 1963, ci-après dénommés **S_63** et les syntagmes équivalents de l'ISO 2018, ci-après dénommés **S_18,** nous avons employé le logiciel *QI Macros,* intégré dans le fichier d'analyse. Cette approche nous permet de comparer la distribution des syntagmes nominaux, adjectivaux et verbaux dans les deux vocabulaires choisis comme points de repère dans la terminologie de l'industrie du plastique.

Afin d'obtenir la répartition de **S_63** et **S_18,** nous avons parcouru les étapes suivantes: sélection de la colonne 2 *Structure morphologique S_63* / colonne 4 *Structure morphologique S_18,* sélection *QI Macros,* sélection *Data&Text Mining,* sélection *Pivot Table Wizard.* Le logiciel nous a offert des résultats surprenants et nous avons décidé de présenter les premières cinq *S_63* / *S_18* et *Total S_63* / *Total S_18:*

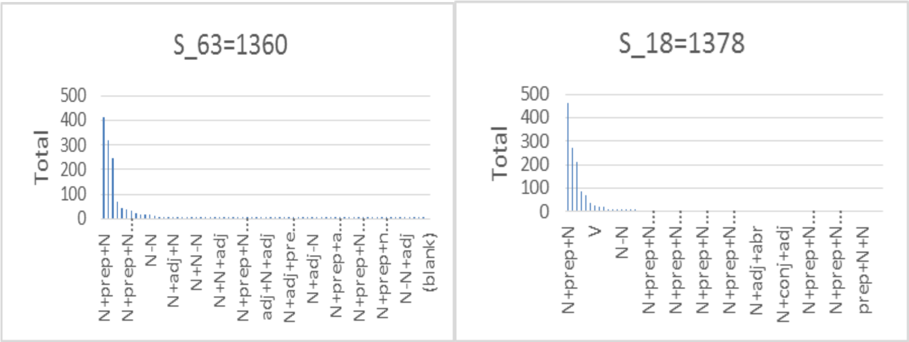
Tableau 5: La répartition des premières cinq S_63 et S_18 par type et nombre d'occurrences

Count of S_63	
S_63[2]	Total
N+prep+N	411
N	319
N+adj	244
V	70
N+prep+N+adj	44
…. **Total S_63**	**1360**

Count of S_18	
S_18[3]	Total
N+prep+N	461
N+adj	274
N	211
N+prep+N+prep+N	85
N+prep+N+adj	72
…. **Total S_18**	**1378**

Les résultats obtenus à partir de mesures statistiques (mesures réalisées en suivant les étapes: sélection de la colonne2 *Structure morphologique S_63* / colonne 4 *Structure morphologique S_18,* sélection *QI Macros,* sélection *Pareto Chart*) sont illustrés par l'intermédiaire de graphiques suivantes:

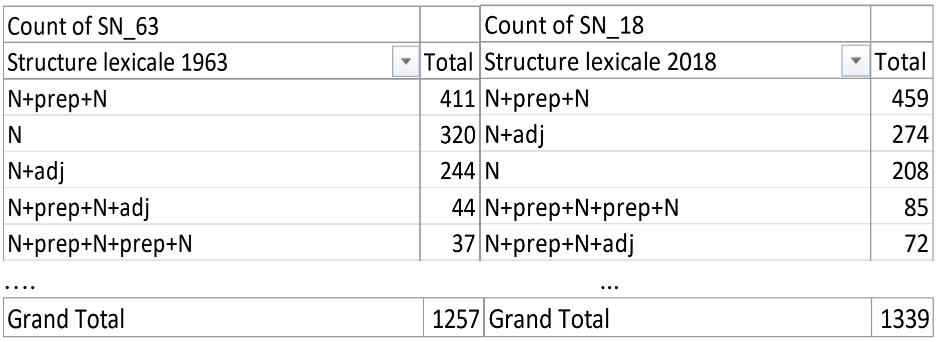
Figure 2: La distribution de S_63 et S_18 par type et nombre d'occurrences



4.1 Les syntagmes nominaux (SN_63 < SN_18)

Afin d'obtenir la répartition de **SN_63** et **SN_18,** nous avons suivi les étapes suivantes: sélection de la colonne 2 *Structure morphologique S_63* / colonne 4 *Structure morphologique S_18,* sélection *Sort&Filter-Filter,* sélection *syntagmes contenant N (nom*), sélection *QI Macros,* sélection *Data&Text Mining,* sélection *Pivot Table Wizard.* Le logiciel nous a offert des résultats surprenants et nous avons décidé de présenter les premières cinq *SN_63* / *SN_18* et *Total SN_63* / *Total SN_18:*

Tableau 6: La répartition des premières cinq SN_63 et SN_18 par type et nombre d'occurrences

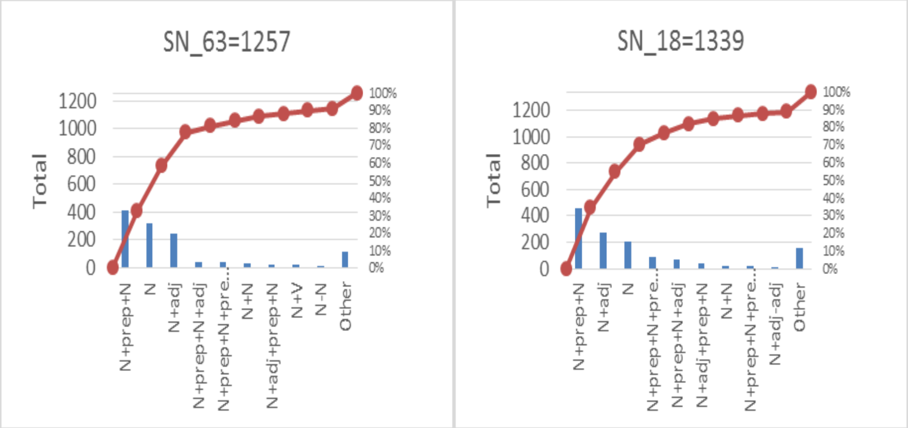


Count of SN_63		Count of SN_18	
Structure lexicale 1963	Total	Structure lexicale 2018	Total
N+prep+N	411	N+prep+N	459
N	320	N+adj	274
N+adj	244	N	208
N+prep+N+adj	44	N+prep+N+prep+N	85
N+prep+N+prep+N	37	N+prep+N+adj	72
….		...	
Grand Total	1257	Grand Total	1339

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1942 (Article)

Les résultats obtenus à partir de mesures statistiques (réalisées en suivant les étapes: sélection de la colonne 2 *Structure morphologique S_63* colonne 4 *Structure morphologique S_18,* sélection *QI Macros,* sélection *Sort&Filter-Filter,* sélection syntagmes nominaux, sélection *QI Macros,* sélection *Data&Text Mining*, sélection *Pivot Table Wizard,* sélection *Pareto Chart*) sont illustrés par l'intermédiaire de graphiques suivantes:

Figure 3: Distribution de SN_63 et SN_18 par type et nombre d'occurrences



Les syntagmes nominaux comprennent **1257** occurrences en 1963 *(SN_63*[4] *= 1257),* où le *Nom* constitue *le terme de base* auquel s'ajoute plusieurs mots (Pitar 2018: 57). Les SN_63 comprennent 21 types de structures, parmi lesquelles les plus répandues sont *N+prep+N* avec 411 occurrences (*adhésif en feuille, assemblage par vis, boudineuse pour câbles, canal d'injection, fil de soudure* etc.). Les prépositions contenues dans cette structure offrent des valeurs sémantiques différentes pour le deuxième terme qui caractérise le premier (Pitar 2018: 57; Bertels 2009):

— partie composante essentielle: *buse* *de* *carotte;*
— le but: *baguette* *d'apport;*
— la caractéristique: *agent* *de* *durcissement.*

La deuxième place est occupée par le nom *(N)* avec 320 occurrences (*baguette, carotte, distributeur, extrudeuse, insertion* etc.) et la troisième place par la structure morphologique *N+adj* avec 244 occurrences *(charge électrostatique, cisaille volante, durcissement excessif, injection capillaire, manchon conique, pièce mauvaise, etc.).*

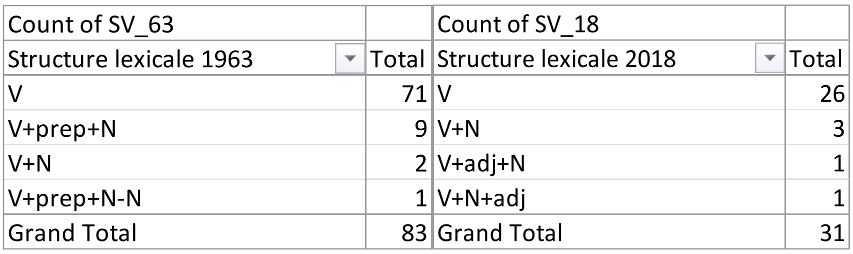
En 2018 nous remarquons **1339 occurrences** (*SN_18*[5] *= 1339),* 21 types de structures parmi lesquelles les plus répandues sont *N+prep+N*avec 461 occurrences (*agent de pontage, adhésif en film, capacité de malaxage, degré de biodégradation, essai en flexion, plaque de démoulage* etc.). Sur la deuxième position on retrouve la structure *N+adj* avec 274 occurrences (*agent anti adhérent, adhésif thermofusible, bande chauffante, capacité physique, caractère thermoplastique* etc.).

La troisième place est occupée par le nom *(N)* avec 208 occurrences qui peut être classifié comme il suit: (a) des noms appartenant à la langue générale (*mélange, forme, renforcement, circulation, alimentation, filtre* etc.) (b) des noms semi-techniques (*mécanisme, plaque, tube, cylindre, filtre* etc.) (c) des noms spécifiques pour le domaine de la plasturgie (*moule, démoulage, bavure, plasticité, thermoformage* etc. (d) des noms qui entrent dans la composition des structures plus complexes, des structures qui désignent des substances chimiques utiles à la fabrication des pièces (*agent – agent de pontage, agent anti adhérent, agent de durcissement, agent bloqué, agent inhibé, agent d'expansion, agent chimique* etc.), des structures qui désignent des procédées de fabrication (*application d'un agent, application de charge, application en construction, application de chaleur, applications électriques, application de l'adhésif, application d'un enduit, applications d'emballage* etc.) ou décrivent différents composants ou outillages (*broche à prisonnier, broche de la compression de l'éprouvette, broche de la complaisance de la machine, broche de chargement* etc.).

4.2 Les syntagmes verbaux (SV_63 > SV_18)

Afin d'obtenir la répartition de **SV_63** et **SV_18**, nous avons suivi les étapes suivantes: sélection de la colonne 2 *Structure morphologique S_63* / colonne 4 *Structure morphologique S_18*, sélection *Sort&Filter-Filter*, sélection syntagmes verbaux, sélection *QI Macros*, sélection *Data&Text Mining*, sélection *Pivot Table Wizard.* Le logiciel nous a offert des résultats surprenants et nous avons décidé de présenter les premières cinq *SV_63* / *SV_18* et le *Total SV_63* / *Total SV_18:*

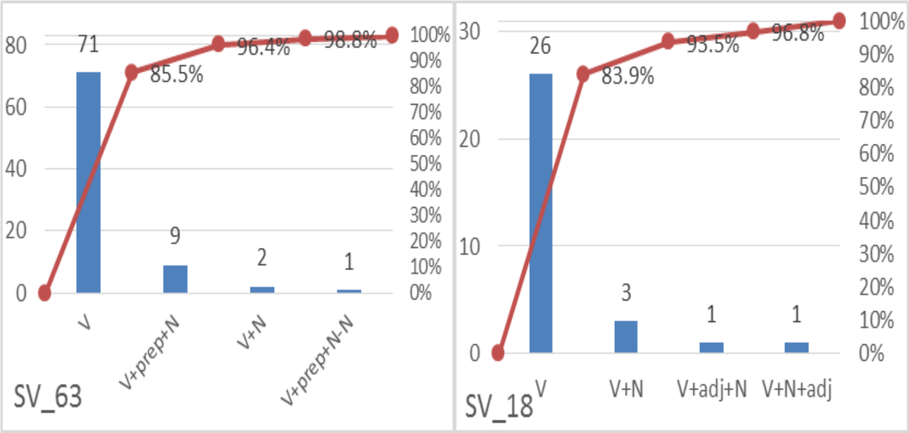
Tableau 7: La répartition des premières cinq SV_63 et SV_18 par type et nombre d'occurrences



Count of SV_63		Count of SV_18	
Structure lexicale 1963	Total	Structure lexicale 2018	Total
V	71	V	26
V+prep+N	9	V+N	3
V+N	2	V+adj+N	1
V+prep+N-N	1	V+N+adj	1
Grand Total	83	Grand Total	31

Les résultats obtenus à partir de mesures statistiques (réalisées en suivant les étapes: sélection de la colonne 2 *Structure morphologique S_63* / colonne 4 *Structure morphologique S_18*, sélection *QI Macros*, sélection *Sort&Filter-Filter*, sélection syntagmes verbaux, sélection *QI Macros*, sélection *Data&Text Mining*, sélection *Pivot Table Wizard*, sélection *Pareto Chart*) sont illustrés par l'intermédiaire de graphiques suivantes:

Figure 4: Distribution de SV_63 et SV_18 par type et nombre d'occurrences (SV_63 > SV_18)



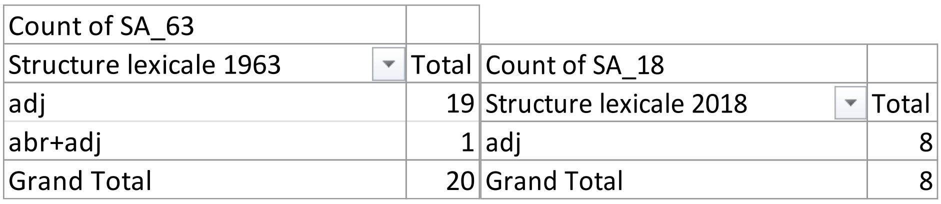
En 1963 nous avons obtenu **83 occurrences.** La première position est occupée par le verbe (V), avec 71 occurrences. Celui-ci peut être systématisé comme suit: (a) des verbes de la langue générale: *abaisser, circuler, adhérer, allonger, cisailler, coller* etc.; (b) des verbes semi-techniques: *calibrer, comprimer, cylindrer, dégazer* etc.; (c) des verbes techniques spécifiques pour le domaine de la plasturgie: *cisailler (la carotte), ébarber, ébavurer, écarotter, éjecter, extruder, injecter, mouler,* etc.; (d) des verbes qui entrent dans la composition des structures plus complexes, des structures qui désignent des procédées de fabrication (*appliquer – appliquer à la spatule, appliquer au pistolet, appliquer en tonneau, faire respirer le moule, enduire par extrusion-laminage*) ou décrivent différents outillages *(châssis d'une machine à former, machine à cisailler, machine à égaliser, machine à enduire, machine à enduire a la brosse, machine à enduire avec lame d'air, machine à enduire avec racle sur rouleau, machine à enduire avec rouleaux inverses, machine à enduire sur rouleaux, machine à imprégner et à enduire, matière à mouler etc.).* La deuxième place est occupée par la structure *V+prep+N* avec 9 occurrences (*enduire à la racle, enduire avec cylindres, enduire au tambour, enduire par extrusion- laminage, reboucher avec de l'enduit etc.)* et la troisième place la structure *V+N* avec 2 occurrences (*couper les bords, faire respirer le moule).*

En 2018 seulement **31 occurrences** et 4 types de structures sont retrouvées dans le vocabulaire étudié, avec 26 verbes sur la première place (*éjecter, séparer, traiter, transformer, travailler* etc.), 3 structures de type *V+N* (*appliquer une protection, appliquer un solvant* et *appliquer la feuille)* et 1 structure de type *V+adj+N (appliquer une mince couche).*

4.3 Les syntagmes adjectivaux (SA_63[6] > SA_18[7])

Afin d'obtenir la répartition de **SA_63** et **SA_18**, nous avons suivi les étapes suivantes: sélection de la colonne 2 *Structure morphologique S_63* / colonne 4 *Structure morphologique S_18*, sélection *Sort&Filter-Filter*, sélection syntagmes adjectivaux, sélection *QI Macros*, sélection *Data&Text Mining*, sélection *Pivot Table Wizard*. Le logiciel nous a offert des résultats surprenants et nous avons décidé de présenter les *SA_63* / *SA_18* et le *Total SA_63* / *Total SA_18:*

Tableau 8: La répartition de SA_63 et SA_18 par type et nombre d'occurrences

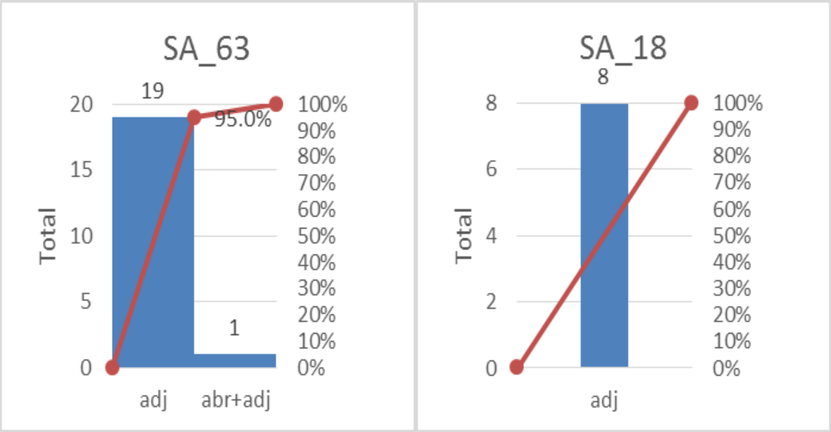


Count of SA_63	
Structure lexicale 1963	Total
adj	19
abr+adj	1
Grand Total	20

Count of SA_18	
Structure lexicale 2018	Total
adj	8
Grand Total	8

Les résultats obtenus à partir de mesures statistiques (réalisées en suivant les étapes: sélection de la colonne 2 *Structure 1963* / colonne 4 *Structure 2018*, sélection *QI Macros*, sélection *Sort&Filter-Filter*, sélection syntagmes adjectivaux, sélection *QI Macros*, sélection *Data&Text Mining*, sélection *Pivot Table Wizard*, sélection *Pareto Chart*) sont illustrés par l'intermédiaire de graphiques suivantes:

Figure 5: Distribution de SA_63 et SA_18 par type et nombre d'occurrences



En 1963 le logiciel a identifié **20 occurrences** (flottant, inamovible, mou, mouillant, moulé, plastifié, pliable, préimprégné, pressée) et en 2018 nous avons identifié **8 occurrences** de type adjectif (fixe, lié, plastifié, thermoplastique, pressée, thermodurcissable). Il faut souligner qu'un nombre important d'adjectifs jouent un rôle crucial dans la formation des structures nominaux (SN) et qu'ils n'ont pas été inclus dans la catégorie SA, mais dans la catégorie SN.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1942 (Article)

5. La productivité des termes candidats (*TC*)

Après la présentation de la situation des structures nominales, verbales et adjectivales, il est absolument indispensable d'analyser la productivité de certains termes candidats indexés dans l'ouvrage rédigé en 1963 et retrouvés sous la même forme ou retrouvés dans des structures nouvelles. Nous avons procédé à l'identification des structures communes en réalisant une comparaison entre la colonne 1 *Lexicon 1963* et la colonne 3 *ISO 2018Vocabulaire plastique*. Nous avons confronté les deux vocabulaires et nous avons extrait manuellement les termes communs et à la suite du filtrage manuel, le classement des mots par l'ordre des structures formés en 2018 nous permet de distinguer certaines caractéristiques parmi les *TC*. Il reste à noter que **120** *TC* ont été classés comme termes qui apparaissent dans les deux vocabulaires et que leur diversité en termes de productivité n'est pas surprenante. Les *TC* retrouvés ont été classés du point de vue de la productivité en trois catégories. *Les TC caractérisés par une forte productivité (N_S63 < N_S18)* sont les plus répandus, et ils comprennent des termes qui ont formé de nouvelles structures au fil du temps. *Les TC caractérisés par symétrie (N_S63 = N_S18)* sont moins répandus et ils décrivent les *TC* qui nous offrent à première vue une symétrie entre les deux lexiques, avec un nombre égal de structures en 1963 et 2018. *Les TC caractérisés par involution (N_S63 > N_S18)* n'ont pas évolué ou ils n'apparaissent plus dans le ISO 2018. Leur évolution ou involution peut être illustrées comme il suit:

Tableau 9: Classification des *TC*

	N_TC[8]	*Proportion par rapport au NT_TC*
Forte productivité	*70*	*58,33%*
Symétrie	*4*	*3,33%*
Involution	*46*	*38,33%*

5.1 Classification des *TC* selon leur productivité

5.1.1 *Des TC caractérisés par une forte productivité (N_S63 < N_S18)*

Cette catégorie comprenne **70** *TC*, ça veut dire **58,33%** du *NT_TC*[9], des unités lexicales qui ont connu une croissance en ce qui concerne les structures produites et indexées dans l'ISO 2018. Dans ce cadre nous allons signaler des structures simples qui ont évoluées dans des structures plus complexes, par exemple le terme *plastique* qui se retrouve dans **7** structures dans les années 1960, *plastique expansé, plastique moulé, plastique rigide, plastique alkydes, plastiques renforcés, plastique à base de caséine, plastique renforcé à la fibre de verre* et qui en **2018** entre dans la composition de **61** structures classées de la manière suivante:

(a) ***plastique + adjectif qui désigne des substances chimiques**: plastique acrylique, plastique acrylonitrile-butadiène-styrène (plastique ABS), plastique acrylonitrile-méthacrylate de méthyle (plastique AMMA), plastique (éthylène/propylène) perfluoré, plastique poly(chlorure de vinylidène), plastique polyallyle, etc.*

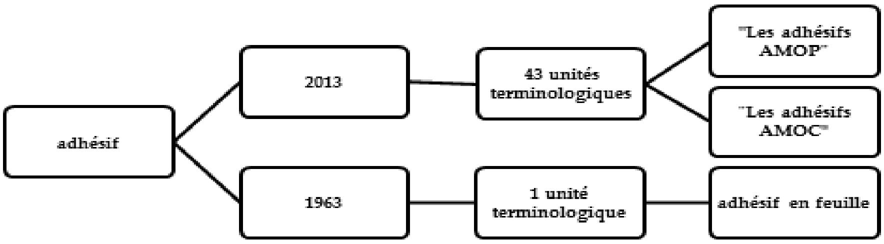
(b) ***plastique + adjectifs**: plastique mélangé, plastique dégradable, plastiques stratifiés, plastique recyclé, plastique retraité, plastique réutilisé, plastique semi-rigide, plastique rigide, plastique vierge, plastique alvéolaire, plastique cellulosique, plastique alvéolaire syntactique, élastomère thermoplastique etc.*

(c) ***plastique + différentes combinaisons adjectif, nom, préposition pour désigner le procédé utilisé pour fabriquer ce type de plastique**: plastique expansé chimiquement, plastique expansé mécaniquement, plastique dégradable par oxydation, plastique expansé thermiquement.*

Les structures *plastique expansé* et *plastique rigide* recensées dans le Lexicon du 1963 sont les seules proposées par l'ISO 2018 avec des structures nouvelles *plastique rigide – plastique semi-rigide, plastique expansé – plastique expansé chimiquement, plastique expansé mécaniquement, plastique expansé thermiquement,* qui permettent la dénomination et la description de nouvelles méthodes de fabrication du plastique.

Un autre terme candidat qui est caractérisé par une forte productivité est le terme *adhésif*[10]. Le «phénomène de factorisation des connaissances» (Lethuillier 2003: 389) peut être identifié en analysant l'arborescence du terme *adhésif*. Le rapport entre les deux lexiques illustre clairement le manque de symétrie entre les deux. Si nous prenons le cas du terme *adhésif,* il suffit d'analyser la multitude de types d'adhésif apparus: dans les années 1963 *l'adhésif* apparaît sous une seule forme, *adhésif en feuille,* mais dans les années 2018 ce terme prend encore 42 formes, *adhésif séché, adhésif sensible, adhésif anaérobie, adhésif à durcissement à froid, adhésif à prise à température ambiante etc.* Les structures logiques présentées sont construites autour du terme *adhésif,* qui peut être nommé racine (Zufferey et Moescheler 2015), par la dérivation par composition, pendant que des autres termes ont été ajoutés afin de signaler la distinction entre plusieurs catégories des adhésifs (*adhésif conducteur, de contact, en émulsion* etc.).

Figure 6: L'évolution du terme adhésif



Il est essentiel de noter que dans les 1960 il n'existait pas une catégorisation des adhésifs puisque les connaissances chimiques avec des influences dans le domaine de la plasturgie n'étaient pas suffisamment étendues pour offrir une variation des adhésifs. L'essor technique se reflète ainsi dans la figure 6 et nous montre que la terminologie de la plasturgie se distingue tout d'abord par la flexibilité en établissant et en construisant des unités terminologiques pour exprimer et désigner les nouvelles réalités de ce domaine[11].

Une vue synthétique des *TC caractérisés par une forte productivité (N_S63 < N_S18)* obtenue pour les *S63* et *S18* est proposée dans le tableau 10, où les *TC* comme *adhésif, essai, plastique, masse* sont des exemples de ce type dans la terminologie de la plasturgie. *Le nombre total de structures (NT_S)* est calculé en réalisant l'addition du *Nombre de structures en 1963* et *Nombre de structures en 2018 (N_S63* et *N_S18).* La productivité des premiers dix *TC* est calculée selon le *N_S63, N_S18* et *NT_S* et le tableau montre une nette variation en ce qui concerne la *proportion de S63 par rapport au NT_S* et la *proportion de S18 par rapport au NT_S.* On peut supposer que leur signification est moins spécifique dans les années 1960, en revanche, les *TC* deviennent plus spécifiques au fil du temps, parallèlement avec les innovations dans le domaine de la chimie et la plasturgie et dans ce cas nous pouvons mentionner le *TC adhésif* sur la première position avec 1 *S63 (adhésif en feuille),* 42 *S18 (adhésif séché, adhésif sensible, adhésif anaérobie, adhésif à durcissement à froid* etc.), *une productivité TC* de 42 et *une proportion de S18 par rapport au NT_S* de 97,67%. Le *TC essai* occupe la deuxième place avec *une productivité* de 11 et *une proportion de S18 par rapport au NT_S* de 91,67%, suivi par le *TC plastique,* avec *une productivité* de 8,71 et *une proportion de S18 par rapport au NT_S* de 89,70%. Les *TC masse* et *temps* occupent la quatrième et la cinquième position avec *une productivité* de 3,6 et 3 et *une proportion de S18 par rapport au NT_S* de 78,27% et 75%. La comparaison entre la productivité des *TC* présente des proportions qui varient beaucoup, entre le premier *TC* et les autres *TC,* avec une diminution significative entre le premier *TC,* qui a connu une productivité de 42 et les *TC* situés sur la deuxième et troisième position, qui comportent une productivité de 11 et 8,76. Mais la comparaison entre les derniers sept *TC* nous montre des tendances identiques, avec une productivité stable située entre 3,6 et 1,34. La productivité du *TC adhésif* s'oppose à la production la plus faible du *TC moulage.*

Tableau 10: La productivité des premières dix *TC*

Terme candidat (TC)	*Nombre de structures en 1963 (N_S63)*	*Nombre de structures en 2018 (N_S18)*	*Nombre total de structures (NT_S = N_S63+N_S18)*	*Productivité TC*[12]	*Proportion de S63 par rapport au NT_S*	*Proportion de S18 par rapport au NT_S*
adhésif	1	42	43	42	2,33%	97,67%
essai	3	33	36	11	8,33%	91,67%
plastique	7	61	68	8,71	10,30%	89,70%

masse	5	18	23	3,6	21,73%	78,27%
temps	6	18	24	3	25%	75%
résine	11	26	37	2,36	29,73%	70,27%
surface	9	20	29	2,22	31,03%	68,97%
résistance	15	33	48	2,2	31,25%	68,75%
procédé	10	20	30	2	33,33%	66,67%
moulage	26	35	61	1,34	42,63%	57,37%

5.1.2 *Des TC caractérisés par symétrie (N_S63 = N_S18)*

La deuxième catégorie comprend des *TC* où le *N_S63 = N_S18*, ça veut dire décrit les *TC* qui nous offrent à première vue une symétrie entre les deux lexiques, avec un nombre égal de structures en 1963 et 2018. Dans cette situation nous avons repéré, après le filtrage manuel, **4** *TC*: *joint, extrusion, assemblage, usiner,* représentés dans le tableau suivant. À l'issue de notre analyse, nous constatons que les *TC* repérés sont caractérisés par symétrie, mais il faut rappeler ici que les structures identifiées dans le lexique du 1963 et 2018 ne possèdent pas la même forme, *S_63 = S_18 (S_63 -assemblage, assemblage par vis* et *S_18 -temps d'assemblage, rupture d'assemblage etc.),* exceptions les *S usiner* et *joint de recouvrement*, qui se retrouvent sous la même forme.

Tableau 11: Les *TC* caractérisés par symétrie

Terme candidat (TC)	*Nombre de structures en 1963 (N_S63)*	*Nombre de structures en 2018 (N_S18)*	*Nombre total de structures — excepté les S communes, (NT_S = N_S63+N_S18)*
joint	11	11	21
extrusion	6	6	12
assemblage	2	2	4
usiner	1	1	1

5.1.3 *Des TC caractérisés par involution (N_S63 > N_S18)*

La troisième catégorie comprend **46** termes candidats, ça veut dire **38.33%** du *NT_TC,* des unités lexicales qui ont connu une décroissance en ce qui concerne les structures produites et indexées dans l'ISO 2018. La vue synthétique de premiers dix *TC caractérisés par involution* nous révèle des *TC* comme: *vis, plaque, machine, moule, manchon, raccord, usinage* etc.

Sur la première position se situent les *TC raccord, manchon, mélangeur, séchoir, flan, filet et vanne qui* ont souffert une réduction de 100% entre les années 1960 et 2018, avec *12,11,11,8, 7,5,3 N_ S63* et 0 *N_S18.* Cette diminution est probablement causée par une réduction concernant le degré d'utilisation des noms de composants de divers outillages et machines, des dénominations qui appartiennent aussi au domaine technique général. Le *TC presse* occupe la deuxième position

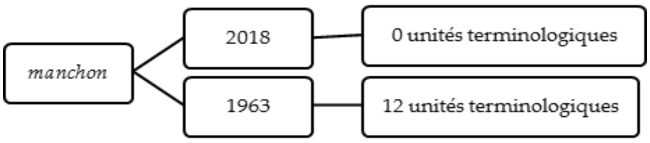
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1942 (Article)

avec une décroissance de 84, 61% et 26 *N_ S63* et *4 N_S18,* suivi par le *TC soudure,* avec une diminution de81,81%, *11 N_ S63* et *2 N_S18.* Dans ce cadre nous allons signaler des *TC* qui n'ont pas évolué et qui n'ont pas été «complétés par de nouveaux éléments» (Halyan 2014: 41), par exemple le *TC manchon* qui se retrouve dans **12** structures dans les années **1960**, *manchon calibré, manchon conique, manchon conique double à rainures internes en dents de scie, manchon de vis, manchon double, manchon double à vis ou visse, manchon élargi, manchon fileté, manchon fretté, manchon extensible etc.* et qui en **2018** n'est plus présent, le cas d'une symétrie n'existant pas entre les deux vocabulaires.

Tableau 12: Les premiers dix *TC* caractérisés par involution

Terme candidat (TC)	*Nombre de structures en 1963 (N_S63)*	*Nombre de structures en 2018 (N_S18)*	*Nombre total de structures (NT_S = N_S63+N_S18)*	*Involution du TC*	*Proportion de S63 par rapport au NT_S*	*Proportion de S18 par rapport au NT_S*
manchon	12	0	12	100%	100%	0%
mélangeur	11	0	11	100%	100%	0%
raccord	11	0	11	100%	100%	0%
séchoir	8	0	8	100%	100%	0%
flan	7	0	7	100%	100%	0%
filet	5	0	5	100%	100%	0%
vanne	3	0	3	100%	100%	0%
presse	26	4	30	84,61%	86,63%	13,37%
soudure	11	2	13	81,81%	84,62%	15,38%
vis	10	2	12	80%	83,34%	16,66%
plaque	39	10	49	74, 35%	79,60%	20,40%
moule	32	10	42	68,75%	76,20%	23,80%
machine	13	5	18	61,53%	72,22%	27,73%

Figure 7: L'involution du *TC* manchon



6. Résultats

Le fichier d'analyse, réalisé à partir du Lexicon 1963 et ISO 2018, nous offre un aperçu rapide des unités lexicales les plus répandues à l'intérieur du secteur de la plasturgie entre les années 1960 et 2018 et il est bien évident que la classification des unités lexicales en trois catégories, *syntagmes nominaux, syntagmes verbaux, syntagmes adjectivaux*, nous fournit une vision globale du vocabulaire utilisé dans cette période. Cabré (1999) mentionne comme caractéristique des langues de spécialité la prédominance des noms et des syntagmes nominaux, qui se retrouvent dans une plus grande proportion que les verbes ou les adjectifs, les résultats illustrés dans le tableau 13 confirmant cette caractéristique:

Tableau 13: Répartition de syntagmes nominaux, adjectivaux et verbaux dans le Lexicon 1963 et l'ISO 2018

Nombre total S_63 (NT_S63[13] = N_SN_63+N_SV_63+N_SA_63)	*Nombre syntagmes nominaux 1963 (N_SN_63)*	*Nombres syntagmes verbaux 1963 (N_SV_63)*	*Nombre syntagmes adjectivaux 1963 (N_SA_63)*	*Proportion de SN_63 par rapport au NT_S63*	*Proportion de SV_63 par rapport au NT_S63*	*Proportion de SA_63 par rapport au NT_S63*
1360	1257	83	20	**92,42%**	**6,10%**	**1,47%**
Nombre total S_18 (NT_S18[14] = N_SN_18+N_SV_18+N_SA_18)	*Nombre syntagmes nominaux 2018 (N_SN_18)*	*Nombres syntagmes verbaux 2018 (N_SV_18)*	*Nombre syntagmes adjectivaux 2018 (N_SA_18)*	*Proportion de SN_18 par rapport au NT_S18*	*Proportion de SV_18 par rapport au NT_S18*	*Proportion de SA_18 par rapport au NT_S18*
1378	1339	31	8	**97,16%**	**2,24%**	**0,58%**

Le tableau antérieur nous permet d'observer des différences entre le vocabulaire de 1963 et les unités retrouvées en 2018 et nous allons attirer l'attention sur le fait que les syntagmes nominaux, qui accomplissent un rôle d'importance indéniable, ont connu une croissance de 4,74%, tandis que les syntagmes verbaux ont connu une diminution de 3,86% et les syntagmes adjectivaux une diminution de presque 1%. Les unités terminologiques indexées dans le document de 1963 et la liste des termes retrouvés ou non dans l'ISO 2018 sont réunies dans le fichier d'analyse, qui a constitué à la fois objet du traitement et de l'examen avec le logiciel *QI Macros*. La terminologie recensée dans le Lexicon de 1963 et recherchée dans l'ISO 2018 contient 1360 termes en 1963 et 1378 termes équivalents en 2018, avec une croissance de 1,32% dans une période de 55 ans.

Nous avons comparé les unités terminologiques recensées dans le Lexicon de 1963 avec le vocabulaire présenté dans l'ISO 2018 et nous avons remarqué que plusieurs structures indexées dans les années 1960 n'apparaissent plus dans le ISO 2018. Il est nécessaire de mentionner que même si les structures ne sont pas recensées dans l'ISO 2018, elles peuvent être toujours utilisées dans les milieux professionnels. À la suite de cette comparaison, les résultats obtenus nous ont permis d'identifier trois catégories de structures qui ne se retrouvent plus parmi les unités terminologiques indexées en 2018:

(a) *des structures faisant partie du vocabulaire général* ont disparu du ISO 2018 puisque l'industrie du plastique a évolué énormément et par conséquent elle a besoin de structures plus spécifiques, plus techniques, par exemple *allonger, balancier, bifurcation, branchement, cisailler, comprimer, convergence, dérivation, doser, embobiner, flasque, inamovible, mouillant* etc.;

(b) *des structures qui désignent des composants, des éléments, des outillages ou parties des outillages* qui ne sont plus d'actualité ou qui sont très connues entre les spécialistes qui opèrent dans l'industrie de la plasturgie, par exemple *manchon, lunette, cadre, mélangeur, garniture, bande support, boudineuse, boudineuse pour câbles, boudineuse à deux vis, broche de guidage, calandre d'enduction, attache de tuyau à un seul filet, bague calibrée, châssis de matrice, filet arrondi* etc.;

(c) *des structures qui sont propres au domaine de la* plasturgie et *qui désignent des opérations, des produits, des procédées de fabrication* etc. qui ne constituent pas de nouveautés ou qui sont moins utilisées ou presque jamais, par exemple *accroche-carotte, accroche-sécheuse, cisailler (la carotte), côté éjection ou fermeture, côté buse ou injection, course de fermeture, course d'ouverture du moule, éjecteur de carotte, formage de feuilles, injection capillaire, méthode de fabrication d'une mousse* etc.

7. Conclusions

Dans notre étude, nous avons réalisé des travaux terminologiques ciblés en nous appuyant sur l'analyse quantitative du lexique de la plasturgie indexé dans les années 1960 et le lexique du ISO 2018. Les mesures statistiques réalisées à l'aide du logiciel *QI Macros* nous ont permis de mettre en valeur les modifications au sein du système terminologique du plastique (Halyan 2014: 43) et d'identifier un ensemble de caractéristiques au niveau morphologique:

— Les unités terminologiques recensées dans le fichier d'analyse ont été systématisées et évaluées selon leur structure morphologique et leur origine: syntagmes nominaux (*SN_63* et *SN_18),* syntagmes verbaux (*SV_63* et *SV_18*) et syntagmes adjectivaux (*SA_63* et *SA_18*). La croissance du nombre de syntagmes nominaux peut être aisément expliquée par la nécessité de la nomination de nouvelles techniques, de nouveaux types de plastique, produits, outillages, procédés de fabrication etc., tandis que la diminution des syntagmes verbaux et des syntagmes adjectivaux peut être justifiée par l'élimination des structures qui renvoient et qui décrivent des procédées de fabrication plutôt dépassés, moins utilisés et des caractéristiques générales, considérées moins spécifiques par les spécialistes du secteur de la plasturgie (*usiner, comprimer, cisailler, flasque, mouillant*).

— Les analyses comparatives effectuées nous ont permis de distinguer trois catégories de termes candidats du point de vue de la productivité: *les TC caractérisés par une forte productivité (N_S63 < N_S18), les TC caractérisés par symétrie (N_S63 = N_S18)* et *les TC caractérisés par involution (N_S63 > N_S18).*

— La plupart des termes, qui sont caractérisés par une forte productivité et qui forment des structures plus complexes dans les années 2010 (*plastique,*

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1942 (Article)

adhésif, masse, temps, résine, surface, résistance, procédé, moulage etc.), sont spécifiques pour le secteur de la plasturgie ou ils appartiennent aux domaines comme la physique et la gestion de la production. En outre, l'étude menée montre que les *TC* qui sont caractérisés par une forte productivité constituent le point de départ dans l'enrichissement de la terminologie du plastique.

— De plus, certains termes propres au domaine de la plasturgie, qui désignent des opérations, des produits, des procédées de fabrication et des termes appartenant au domaine technique, sont caractérisés par une involution et au fil du temps ils ne seront plus inclus dans le lexique de la plasturgie probablement à cause d'une réduction due à leur degré d'utilisation (*éjecteur de carotte, côté buse ou injection, course de fermeture, manchon, raccord, cylindre, vis, flan, mélangeur, poinçon* etc.).

— Du point de vue diachronique, nous avons remarqué une influence indéniable des termes propres au secteur de la plasturgie dans la constitution de nouvelles structures et une contribution majeure du vocabulaire de la physique dans la formation des structures plus spécifiques, qui reflètent les dernières innovations du secteur de la plasturgie.

— À l'issue de notre étude, il semble utile de souligner que les adjectifs et les syntagmes adjectivaux n'ont pas été mis en valeur et nous considérons que l'initialisation des recherches dans ce sens-là nous permettra de repenser le rôle joué par l'adjectif dans la formation des langues de spécialité.

Notes de fin

1. La traduction appartient à l'auteur
2. SN_63 = syntagme nominal indexé dans le Lexicon 1963
3. SN_18 = syntagme nominal équivalent dans l'ISO 2018
4. SV_63 = syntagme verbal indexé dans le Lexicon 1963
5. SV_18 = syntagme verbal équivalent dans l'ISO 2018
6. SA_63 = syntagme adjectival indexé dans le Lexicon 1963
7. SA_18 = syntagme adjectival équivalent dans l'ISO 2018
8. *N_TC* = nombre termes candidats
9. *NT_TC* = nombre total de termes candidats (120)
10. Les adhésifs sont classés, selon *L'Institut Supérieur de Plasturgie d'Alençon,* en 2 catégories: «Les adhésifs de mise en œuvre physique (AMOP) sont des polymères, appliqués sous forme liquide, qui ne subissent qu'un changement d'état physique durant l'étape de collage.» et «Les adhésifs de mise en œuvre chimique (AMOC) sont des adhésifs plus spécifiquement adaptés aux applications structurales et pour lesquels est créé un réseau tridimensionnel à l'issue de la réaction chimique ayant lieu lors du collage (opérations de polymérisation à partir de précurseurs monomères ou oligomères).»

11. À partir des années 1950 les recherches dans le domaine de la plasturgie s'intensifient et au fil du temps de nouveaux types de plastique apparaissent: les thermodurcissables, les élastomères, les adhésifs etc.
12. *Productivité TC* = productivité de termes candidats en analysant les structures du 1963 et les nouvelles structures de 2018 (N_S18 divisé par N_S63)
13. *NT_S63* = Nombre total de syntagmes indexées dans Lexicon 1963
14. *NT_S18* = Nombre total de syntagmes équivalentes dans l'ISO 2018

Références bibliographiques

Arhire, M. 2014. *Corpus-based Translation for Research, Practice and Training.* Iasi: Institutul European.

Avornicesei, O.-F. et al. 2021. *Studii de traducere și terminologie specializată.* Cluj-Napoca: Risoprint.

Bertels, A. 2009. *Etudier la sémantique des termes techniques: des théories à la pratique.* Consulté le 15 janvier 2023: https://www.researchgate.net/publication/228941070

Bidu-Vrânceanu, A. 2000. *Lexic comun, lexic specializat.* Bucarest: Editura Universității București.

Bidu-Vrânceanu, A. 2007. *Lexicul specializat în mișcare de la dicționare la texte.* Bucarest: Editura Universității București.

Cabré, M.T. 1999. *Terminology: Theory, Methods, and Applications.* Amsterdam/Philadelphie: John Benjamins.

Cabré, T. 2016. La terminologie. Forner, W. et B. Thörle. 2016. *Manuel des langues de spécialité:* 68-81. Berlin/Boston: De Gruyter.

Dankova, K. 2021. La néologie terminologique en français dans le domaine des fibres chimiques. *Academic Journal of Modern Philology* 13: 97-108. Consulté le 11 décembre 2022: https://www.ceeol.com/search/article-detail?id=1003127

Forner, W. et B. Thörle. 2016. *Manuel des langues de spécialité.* Berlin/Boston: De Gruyter.

Gaudiaut, T. 2020. 70 ans d'industrie plastique. *Statista* 2020, January 10. Consulté le 3 novembre 2022: https://fr.statista.com/infographie/20457/evolution-de-la-production-mondiale-de-plastique/

Halyan, O. 2014**.** La structure et la sémantique des groupes de mots termes physiques de français. *Romanica Cracoviensia* 1: 38-48. Consulté le 9 novembre 2022: https://www.ceeol.com/search/article-detail?id=428153

Ilinca, E.-C. et A.-M. Tomescu. 2013. Aspects lexico-sémantiques de la traduction technique du français vers le roumain. *Traduire* 228: 68-80. Consulté le 2 mars 2023: http://journals.openedition.org/traduire/534; DOI: https://doi.org/10.4000/traduire.534

Institut Supérieur de Plasturgie d'Alençon. 2014. *Etat de l'art, Méthodes d'assemblage de thermoplastiques renforcés fibres courtes ou longues.* Consulté le 21 octobre 2022: https://www.nae.fr/wp-content/uploads/2018/04/EB_Assemblage_TP.pdf

International Organization for Standardization. (n.d.) *ISO 472:2013(fr) Plastiques — Vocabulaire*. Consulté le 15 octobre 2022: https://www.iso.org/obp/ui/#iso:std:iso:472:ed-4:v1:fr

Jacquey, E., L. Kister, M. Marcon et S. Barreaux. 2018. Termes complexes et langues de spécialité en sciences humaines et sociales: que nous apprennent les textes intégraux? *Revue Meta* 63(1) 7-29. https://doi.org/10.7202/1050512ar

Janssen, M. et M. van Campenhoudt. 2005. Terminologie traductive et représentation des connaissances: l'usage des relations hyponymiques. *Langages* 39(157): 63-79. Consulté le 27 octobre 2022:
doi: https://doi.org/10.3406/lgge.2005.975
https://www.persee.fr/doc/lgge_0458-726x_2005_num_39_157_975

Laroche, A. et al. 2011. Étude de l'influence de la taille du corpus de référence sur l'extraction terminologique automatique contrastive. *Long papers of the 9th International Conference on Terminology and Artificial Intelligence, TIA 2011, Paris, 8–10 November 2011:* 66-72. Consulté le 3 novembre 2022:
http://olst.ling.umontreal.ca/pdf/Laroche_et_al_2011.pdf

Lethuillier, J. 2003. L'enseignement des langues de spécialité comme préparation à la traduction spécialisée. *Revue Meta* 48(3): 379-392. Consulté le 4 janvier 2023:
https://doi.org/10.7202/007598ar

N'Da, P. 2015. *Recherche et méthodologie en sciences sociales et humaines. Réussir sa thèse, son mémoire de master ou professionnel, et son article.* Paris: L'Harmattan.

Pierre, L. et B. Parth. 2005. *Structure du français moderne. Introduction à l'analyse linguistique.* Toronto: Canadian Scholars' Press.

Pitar, M. 2018. *Manual de terminologie și terminografie.* Timișoara: Editura Universității de Vest.

PlasticsEurope. (n.d.) *Les plastiques, des matériaux innovants.* Consulté le 8 février 2023:
https://legacy.plasticseurope.org/fr/about-plastics/what-are-plastics/innovative-material

Polzin-Haumann, C. et W. Schweickard. 2015. *Manuel de linguistique française.* Berlin/Munich/Boston: De Gruyter.

Tomescu, A.-M. 2022. Caractéristiques linguistiques et terminologie spécifique au domaine des voitures électriques. *Studii şi cercetări filologice. Seria Limbi Străine Aplicate* 21: 144-155. Consulté le 18 décembre 2022:
https://www.ceeol.com/search/article-detail?id=1088363

Vidal-Gorène, C. et al. 2020. Modèles d'annotations morphologiques pour le traitement de données multivariées de l'arménien. *2èmes journées scientifiques du Groupement de Recherche Linguistique Informatique Formelle et de Terrain (LIFT), Dec. 2020, Montrouge (virtuel), France: 72-82.* Consulté le 13 octobre 2022:
https://hal.science/hal-03047147

Wittfoht, A.M. et al. 1963. *Plastics Lexicon Processing and Machinery — German–English–French–Spanish–Italian–Dutch.* Amsterdam: Elsevier.

Zanola, M. 2021.Terminologie diachronique: méthodologies et études de cas. Introduction. *Cahiers de lexicologie* 118(1): 13-21.

Zufferey, S. 2020. *Introduction à la linguistique du corpus.* Londres: ISTE Editions.

Zufferey, S. et J. Moescheler. 2015. *L'initiation à la linguistique française.* 2e édition. Paris: Colin Armand.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1944 (Article)

Using Generative AI to Provide High-Quality Lexicographic Assistance to Chinese Learners of English

Qian Li, *Centre for Lexicographical Studies, Guangdong University of Foreign Studies, China (lqchristina@gdufs.edu.cn) (https://orcid.org/0009-0002-8267-7762)*
and
Sven Tarp, *Centre for Lexicographical Studies, Guangdong University of Foreign Studies, China; Department of Afrikaans and Dutch, Stellenbosch University, South Africa; and Aarhus University, Denmark (st@cc.au.dk) (https://orcid.org/0000-0003-1941-9082)*

Abstract: This paper reports on a research project that aims to explore how and to what extent generative AI can be used to produce different types of explanations that can be activated in writing assistants for Chinese learners of English. It first places the project in a lexicographic context and describes the general methodology used, including the limited usefulness of a learner corpus as an empirical basis and the need to use ChatGPT as a supplement to determine the error sub-categories to be explained. As a result, 26 error sub-categories are identified within the main category of subject–verb disagreement. The paper then compares two generative AI chatbots, Baidu's Ernie Bot and OpenAI's ChatGPT, and describes how the latter was found to be more efficient and therefore prompted by lexicographers with experience in second-language teaching to write long explanations for each of the error sub-categories, with several examples demonstrating both the chatbot's remarkable performance and the constant need for human supervision and intervention. At the same time, the paper argues for the integration of generative AI directly into writing assistants to produce short default explanations for errors found in learners' texts. Finally, the paper summarises the findings, including the complex relationship between human and artificial intelligence.

Keywords: AUTOMATIC ERROR CORRECTION, CHATBOTS, ERROR EXPLANATIONS, FREQUENCY CRITERIA, GENERATIVE AI, L2 LEARNING, LANGUAGE MODELS, LEARNER CORPUS, MODERN GLOSSES, WRITING ASSISTANTS

Opsomming: Die gebruik van generatiewe KI om hoëkwaliteit leksikografiese hulp aan Chinese aanleerders van Engels te bied. In hierdie artikel word verslag gelewer oor 'n navorsingsprojek wat daarop gemik is om te ondersoek hoe en tot watter mate generatiewe KI gebruik kan word om verskillende tipes verklarings te verskaf wat in skryfhulpmiddels vir Chinese aanleerders van Engels geaktiveer kan word. Die projek word eerstens in 'n leksikografiese konteks geplaas en die algemene metodologie wat gebruik word, word beskryf. Die beperkte bruikbaarheid van 'n leerderkorpus as empiriese basis en die behoefte aan die gebruik van ChatGPT as 'n hulpmiddel om die foutsubkategorieë wat verklaar moet word te bepaal, word hierby ingesluit. Dit het tot gevolg dat 26 foutsubkategorieë binne die hoofkategorie van onderwerp–werkwoord-kongruensie geïdentifiseer is. Twee generatiewe KI-kletsbotte, Baidu se Ernie Bot en OpenAI se ChatGPT, word dan met mekaar vergelyk, en daar word beskryf hoe laasgenoemde meer doeltreffend bevind is. Daarom is ChatGPT deur leksikograwe met ervaring in tweedetaalonderrig versoek om lang verklarings vir elk van die foutsubkategorieë te skryf, wat verskeie voorbeelde insluit wat beide die kletsbot se merkwaardige werkverrigting en die konstante behoefte aan menslike toesig en intervensie demonstreer. Terselfdertyd word die direkte integrasie van generatiewe KI in skryfhulpmiddels bepleit om kort verstekverklarings vir foute wat in leerders se tekste gevind word, te lewer. Laastens word die bevindings, insluitend die komplekse verhouding tussen menslike en kunsmatige intelligensie, opgesom.

Sleutelwoorde: OUTOMATIESE FOUTKORRIGERING, KLETSBOTTE, FOUTVERKLARINGS, FREKWENSIEKRITERIA, GENERATIEWE KI, L2-LEER, TAALMODELLE, AANLEERDERSKORPUS, MODERNE GLOSSE, SKRYFHULPMIDDELS

1. Introduction

In his reflections on the future of lexicography, and in response to Grefenstette's (1998) famous question of whether there will be lexicographers in the year 3000, Rundell (2012: 18) optimistically predicts that there will still be lexicographers, but that they will be doing something different from what their 21st century colleagues are doing. We agree with Rundell in principle, but we would like to emphasise even more that the lexicographers of the future will not only carry out their work using different methods and techniques from those of today. The results of this work are also likely to be presented to future users in entirely new ways. As McArthur (1986) has shown, lexicography has undergone similar shape-shifting from time to time in its millennia-long evolution from clay tablet to computer, and there is no reason to doubt that it will not do so again in the future. With this in mind, Tarp and Gouws (2023) have proposed a redefinition of the discipline of lexicography to include not only dictionaries but also glosses, both the traditional ones from which dictionaries have evolved according to Hanks (2013) and Benati and Händl (2019), and the new ones that are emerging, supported by cutting-edge technologies and integrated into writing and reading aids as well as other kinds of digital software. Tarp and Gouws (2023: 439) therefore recommend that lexicographers should:

> shift their focus from dictionaries to databases containing both new and old types of lexicographical data that can serve various tools, including but not limited to digital dictionaries.

From this perspective, Tarp and Gouws distinguish between two different categories of lexicographic databases that can already be observed in practice, namely "traditional" *lemma-centred databases* and new *problem-centred databases*. The latter

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1944 (Article)

do not focus on specific words (lemmas), but on classes of grammatical, orthographic and stylistic challenges and problems that appear in texts. Because of these characteristics, problem-centred databases cannot support dictionaries as we know them, as they only contain data (glosses) that can be visualised in various digital tools to explain problems and help solve language challenges. These glosses are not related to specific words, but to specific types of problems, usually associated with a wider group of words.

It is obvious that the preparation, organisation and usefulness of problem-centred databases are much less studied than those of lemma-centred databases. However, it is not just that the latter have been around for longer. It is also a matter of taking a broader view of lexicography and breaking new ground. An example of this is the increasing use of Generative Artificial Intelligence (AI) in the discipline, especially after the launch of OpenAI's ChatGPT in November 2022. To date, most of the academic publications on the subject have focused on the use of this new technology to perform various tasks related to dictionary making; see, for instance, Alonso-Ramos (2023), Jakubíček and Rundell (2023), Lew (2023), Phoodai and Rikk (2023), Rees and Lew (2023), Rundell (2023), De Schryver (2023), and McKean and Fitzgerald (2024). So far, Huete-García and Tarp (2024), Li, Tarp and Nomdedeu-Rull (2024), and Tarp and Nomdedeu-Rull (2024), who are all concerned with the creation of lexicographic data to be used in writing assistants, are among the few exceptions to this trend. And the same can be said of Abdullayeva and Muzaffarovna (2023), Song and Song (2023) and Wu (2024), who approach writing tools from a different disciplinary perspective.

Against this background, we have conducted a research project to explore how and to what extent generative AI can be applied to produce different types of glosses — hereafter referred to as explanations — that can be activated in writing assistants for Chinese learners of English. The hypothesis is that this technology can increase productivity, at least without compromising quality, but probably improving it as well. This hypothesis is based on some reflections made by Huete-García and Tarp (2024), who experimented with ChatGPT to develop a writing assistant for learners of Spanish. The two researchers make a distinction between oral and written communication from teacher to student. On the one hand, they note that experienced Spanish teachers can easily explain the different types of language problems and challenges to their students in class. On the other hand, however, Huete-García and Tarp (2024: 36) observe that:

> it is less straightforward to write a concise explanation that gets to the heart of the matter in a language that is easily understood by the target audience. In addition to selecting the key aspects to be covered, determining the most appropriate and pedagogical structure can be quite time-consuming.

They therefore recommend using ChatGPT for this task, but only as an inspiration, as experienced teachers or lexicographers should always have the last say. Li et al. (2024), who have further developed and tested this way of writing explanations, define it as a "necessary symbiosis" between human and artificial intelli-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1944 (Article)

gence. As lexicographers with experience in second-language teaching, we can easily recognise ourselves in the above description and have therefore adopted the same approach in our project.

In the next section, we will briefly explain the overall methodology used to carry out the project, including why we have based the work on an English corpus containing errors made by Chinese learners of different proficiency levels. Section 3 describes how some of the error types to be explained are determined. Section 4 reports on the main part of the project, i.e. the direct work with generative AI, where two different chatbots are used to generate explanations and their efficiency is compared. Section 5 then summarises the main findings and presents the general conclusions, together with some reflections on future work.

2. Methodology

The lexicographic glosses, i.e. the explanations that are the subject of the research project described here, cannot be planned, produced or evaluated without knowing exactly how they will be used and what specific purpose they will serve. So, the very first step to be taken is to clearly identify and define that purpose, i.e. *who* might need the explanations, *for what* they might need them, *in what situation* they might need them, and *in what technological environment* the need might arise.

Now, the explanations are intended to help Chinese beginner and intermediate learners of English who are writing English texts using an AI-based writing assistant, similar in many ways to Grammarly or ProWritingAid (see Fitria 2021, 2023), but unlike these, bilingual with explanations in Chinese, i.e. the target users' native language. It is trained to identify and highlight possible problems and suggest alternative solutions that Chinese learners may want to understand in more detail as part of their English learning process. *Helping the learners achieve this deeper understanding is the genuine purpose of the explanations*.

Thus, unlike Wiegand's (1987) classic concept of "genuine purpose", which refers to a dictionary as a whole, here it refers only to the explanations, but not to the writing assistant as such. The reason for this is that the assistant has a broader purpose related to the writing process. Apart from simple text correction, the alternative suggestions generated by the underlying language model for this specific purpose are presented in a way that supports *incidental learning*, as defined primarily in relation to reading by Krashen (1989), Shu, Anderson and Zhang (1995) and Hulstijn (2013), among others, and later adapted to writing and even lexicography by Graham (2020) and Tarp (2022), respectively. Finally, as an additional service to motivated learners, the design also allows them to move on to *intentional learning* — see Leow and Zamora (2017) — if they decide to access and read the detailed explanations that are the subject of this paper.

All this suggests that the explanations should be written as *short didactic texts in plain language, without too much technical terminology, providing the most*

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1944 (Article)

relevant information about the specific language problem, and structured in a way that makes it easy for the reader to get an overview and grasp the essence of the problem.

As mentioned in the previous section, writing short didactic texts with these characteristics can be time-consuming, even for experienced and knowledgeable second-language teachers, as it usually requires some prior in-depth reflection on content, style and structure. It might therefore be interesting to explore whether, how and to what extent lexicographers can benefit from generative AI in this task. For this purpose, two well-known chatbots were chosen, namely Baidu's *Ernie Bot* and OpenAI's *ChatGPT*. Both of them were instructed to write explanations of selected problems in both Chinese and English. This means that four different approaches or methods were used to test their performance for this specific purpose, after which their respective performances were compared. We are fully aware that Chinese generative AI chatbots like Ernie Bot are generally considered to be a year or two behind the most advanced Western ones, such as ChatGPT, but as the writing assistant in question is intended to correct errors made by Chinese learners of English, there may appear to be some deviation from this general "rule". In any case, generative AI is a technology that is developing almost exponentially, and much is expected to change in the next few years. For now, the initial hypothesis was that Ernie Bot would be more efficient at writing Chinese than English, and ChatGPT would be more efficient at writing English than Chinese. But regardless of whether this turns out to be true or false, the comparison of the four methods provided evidence to better determine the most advantageous way to produce the explanations using current technology, i.e. either writing them directly in Chinese, or writing them in English and then translating them into Chinese, a process that poses other challenges.

The *Chinese Learner English Corpus* (CLEC) was used to select the types of errors or problems to be explained. This corpus, compiled by Gui and Yang (2003), is currently the only tagged corpus in China that contains errors made by Chinese learners of English. It consists of a just over a million words divided into five parts of about 200,000 words each, according to the learner's proficiency level. It is a relatively small corpus for the specific task, but its size does not differ much from similar tagged corpora in other languages, as they are very time-consuming and costly to produce.

The use of some kind of learner corpus is definitely a must in order to identify typical learner errors that can be used and explained in didactic language tools. These corpora have been used in one way or another to develop numerous writing tools, as discussed by Bestgen and Granger (2011), Paquot (2012), Wanner, Verlinde and Alonso-Ramos (2013), Alonso-Ramos and García-Salido (2019), Frankenberg-García, Lew, Roberts, Rees and Sharma (2019), and Granger and Paquot (2022). The best type of corpus for this purpose is undoubtedly a tagged corpus with parallel correction of the errors detected, such as the Spanish one described by Davidson, Yamada, Fernández-Mira, Carando, Sánchez-Gutiérrez and Sagae (2020). There are two main types of tagged or parallel corpora, namely

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1944 (Article)

those that contain errors made by real learners, and those in which these errors — also referred to as "noisy examples" — are introduced using different types of software, such as those presented by Xie, Genthial, Xie, Ng and Jurafsky (2018) and Zhao, Wang, Shen, Jia and Liu (2019). According to the former, it is now possible to "synthesize noisy examples that human evaluators" are "nearly unable to discriminate from nonsynthesized examples" (Xie et al. 2018: 626). This technique, which can easily generate parallel corpora of several million words, is clearly useful and practical for a whole range of purposes, as also pointed out by Huete-García and Tarp (2024) in relation to their Spanish writing assistant project. In this respect, the "corpus revolution in lexicography" celebrated by Hanks (2012), who himself contributed significantly to its success, has indeed entered a new phase with possibilities and perspectives that have yet to be fully explored.

Be that as it may, a corpus of synthesised "noisy examples" simply does not serve as the basis for writing explanations of the kind discussed in this paper. Learners make a large number of errors of different types, and so does the software that synthesises this "noise" and feeds it into a corpus. Many errors, especially typos and other misspellings, are quite banal and do not lend themselves to explanation. Even if these are eliminated, there will still remain a significant number of error types that will require considerable work and time to explain properly. It is therefore necessary to prioritise, which can be done on the basis of frequency starting with the most common types. To be meaningful, such a frequency determination can only be made from a corpus of human-made errors. Using the above "noise-synthesising" software for this purpose would be arbitrary and the results would not reflect the true frequency of real learners' errors.

This discussion is reminiscent of a similar discussion about frequency as a lemma selection criterion in a traditional dictionary project, best illustrated by Kilgarriff's (2013: 79) idea that "if a dictionary is to have N words in it, they should be the N words from the top of the corpus frequency list". This approach has been challenged by Trap-Jensen, Lorentzen and Sørensen (2014), among others, who argue that corpus frequency is not necessarily identical to look-up frequency. Consequently, Nomdedeu-Rull and Tarp (2024: 174) point to another empirical source, namely log files, which in some cases have recorded hundreds of millions of look-ups in online dictionaries and therefore provide a much more accurate picture of the most frequently consulted words from which the lemmas in a new dictionary project can be selected. In this case, just as in the case above, the challenge is to put the real human users at the centre of the lexicographic work and to focus on their evidence-based needs.

To this end, the *Chinese Learner English Corpus* was used together with the AntConc corpus tool. This allowed us to determine the frequency of the specific error categories relevant to the project once they had been identified. How this was done, as a symbiosis of corpus search, knowledge and generative AI, will be discussed in the next section.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1944 (Article)

3. Determining error categories

The *Chinese Learner English Corpus* groups all identified errors into eleven major domains (*word formation*, *verb phrase*, *noun phrase*, *pronoun*, *adjective phrase*, *adverb*, *preposition*, *conjunction*, *lexical*, *collocational* and *syntactic*). These domains are further divided into 61 general categories, all of them at a very high level of abstraction. As such, they do not lend themselves to explanation in the form of short didactic texts which, as defined above, provide "the most relevant information about the specific language problem". If the 61 categories were explained as they are, such explanations would be either far too long or far too general for learners who only want to know more about a specific problem highlighted by a writing assistant in a text they have written. So, we had to break them down into sub-categories that were more suitable for explanation.

An example of this is *agreement* (or *concord*), which the corpus records as a frequent problem for Chinese learners of English, and which is treated as three different categories under the domains of *verb phrase*, *noun phrase* and *pronoun*, respectively. Each of these categories comprises several sub-categories that are not listed separately in the corpus. Not all of them have the same frequency, but in order to work systematically it was necessary to identify them and then group them under the respective categories, just as Bestgen and Granger (2011: 239-240) did with spelling errors. For the specific purpose of this paper, we chose the overall error category SUBJECT–VERB DISAGREEMENT under the *verb phrase* domain. However, our method differed from theirs in that we decided to experiment with ChatGPT and use it as an inspiration to speed up the sub-categorisation process:

1. First, we asked it to give us a list of relevant error types, which it did surprisingly well.
2. We then clicked the regenerate button to see if it would give us more useful suggestions, which it did in most cases. We kept doing this until it started repeating itself and nothing new came up.
3. If the result was not satisfactory, we also tried modifying or rewriting the prompt to improve the output.
4. The next step was to use our own grammatical knowledge and teaching experience to add a few more error types or to split some of the ones the chatbot provided into two.
5. As proof of the pudding, we consulted the corpus to see if it actually contained errors belonging to all the proposed sub-categories. If it did not, the sub-category was ignored for the time being.
6. Finally, we refined the terms used to describe the different sub-categories, as the chatbot was not consistent in this regard.

The end result was the following list of 26 sub-categories under the overall error category SUBJECT–VERB DISAGREEMENT:

— **Nouns:** Singular noun + plural verb
— **Nouns:** Plural noun + singular verb

- **Compound subjects:** Compound subjects joined by "and" + singular verb
- **Compound subjects:** Compound subjects joined by "or" or "nor" with nearest subject in singular + plural verb
- **Compound subjects:** Compound subjects joined by "or" or "nor" with nearest subject in plural + singular verb
- **Proper nouns:** Title of book, film and other works + plural verb
- **Uncountable nouns:** Uncountable noun + plural verb
- **Uncountable nouns:** Uncountable noun ending in "s" + plural verb
- **Infinitives:** Single infinitive as subject + plural verb
- **Gerunds:** Single gerund as subject + plural verb
- **Gerunds:** Two or more gerunds as subject + singular verb
- **Personal pronouns:** Third-person singular personal pronoun + plural verb
- **Personal pronouns:** Personal pronoun except for third-person singular + singular verb
- **Personal pronouns:** Personal pronoun + verb "to be" inflected in wrong person in present tense
- **Personal pronouns:** Personal pronoun + verb "to be" inflected in wrong person in past tense
- **Indefinite pronouns:** Singular indefinite pronoun + plural verb
- **Indefinite pronouns:** Plural indefinite pronoun + singular verb
- **Demonstrative pronouns:** Singular demonstrative pronoun + plural verb
- **Demonstrative pronouns:** Plural demonstrative pronoun + singular verb
- **Relative pronouns:** Relative pronoun with singular referent + plural verb
- **Relative pronouns:** Relative pronoun with plural referent + singular verb
- **Formal subject "there":** There + singular verb + plural real subject
- **Formal subject "there":** There + plural verb + singular real subject
- **Adverb "here":** Here + singular verb + plural subject
- **Adverb "here":** Here + plural verb + singular subject
- **Additions to subject beginning with "as well as", "together with", "along with", etc.:** Singular subject + addition + plural verb

It is important to note that the above is not a traditional linguistic classification, but one that considers only those cases where it is possible to unambiguously explain the respective sub-categories. In this respect, it should also be noted that the list does not include all the problems we are aware of. *Collective nouns*, for instance, are not included in the list, because, depending on what the writer wants to express, they can be used with both singular (most of the time) and plural verbs, but the technology to distinguish between these two behaviours with reasonable accuracy is not yet available. However, should the language model, after being trained, start to occasionally highlight and suggest alternatives to some verbs in relation to collective nouns, an explanation can be prepared that presents the general rules for this type of agreement, without taking a position on the specific suggestion. And the same can be done in some other cases, such as *subjects made up of two or more infinitives*, which usually require a plural verb, but can also be combined with a singular verb in special circumstances.

Since the described method is not induction from existing empirical data, but a combination of deduction and generative AI, it cannot be excluded that some less frequent sub-categories of the 61 general categories registered in the corpus may be overlooked. However, this should not be a major concern, as they can be easily detected and added when the prototype of the writing assistant is ready to be tested on real users. Although we have no hard facts as such to prove it, we are convinced that the chosen method significantly speeds up the identification process. Once this identification of relevant error sub-categories for our research project was done, we started prompting the two chatbots to write the corresponding explanations. This proved to be a new and unexpected challenge, which will be discussed in the next section.

4. Generating explanations

As mentioned in Section 2, our intention was to find out both whether Baidu's *Ernie Bot* or OpenAI's *ChatGPT* produce the best explanations, and whether these explanations could actually be considered useful for Chinese learners who might need to know more about specific problems identified in one of their texts. Since the ultimate goal is to provide explanations in the learners' native language, both chatbots were instructed to write explanations both directly in Chinese and in English for subsequent translation into Chinese, and the respective results were then compared. To this end, we first designed a default English prompt based on the experience of a similar research project conducted by Li et al. (2024):

> I am training a language model to be used for didactic purposes. I want you to briefly and didactically explain to a Chinese learner of English [problem type]. I need the text to have the following structure 1) an introduction to the problem, 2) an example sentence with an error, 3) the same example sentence with the error corrected and 4) a brief explanation of the incorrect and correct examples. The example sentences must be at least 12 words long.

This prompt was then adapted into Chinese and later modified whenever the respective chatbots' responses were deemed unsatisfactory, either in terms of content, style or terminology used.

4.1 Ernie Bot disappoints

It soon became clear that Ernie Bot had much more difficulty than ChatGPT in producing acceptable explanations, whether in English or Chinese. In some cases, its responses to our prompts contained only a few minor problems that were relatively easy to correct. In other cases, however, it seemed to have completely misunderstood the grammatical issue raised, even though we modified the respective prompts several times. For example, when asked to write an expla-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1944 (Article)

nation of the relatively simple error of using a plural verb together with a singular noun as the subject, it first gave the "incorrect" example: *The boy loves playing footballs and basketballs*. It then corrected this example to: *The boy loves playing football and basketball*, explaining it as follows:

> In the incorrect example, 'The boy loves playing footballs and basketballs', the subject 'The boy' is singular, but the verb 'loves playing' is correctly paired with a singular subject. However, the problem lies in the objects of the verb, 'footballs' and 'basketballs'. Both of these nouns are incorrectly pluralized, as the sports 'football' and 'basketball' are typically treated as singular concepts when referring to the game or sport itself, not individual balls. But more importantly for this exercise, the focus is on the subject–verb agreement. Even if the objects were correctly singular, the subject–verb agreement is not the issue here. The critical point is to emphasize that a singular subject should not be mistakenly thought to require a plural verb.

So, despite being instructed to deal with problems of subject–verb agreement, Ernie Bot first gives two examples of correct agreement in the singular (i.e. no incorrect example of this problem), then claims that "the problem lies in the objects of the verb" because "the subject–verb agreement is not the issue here", and finally states that the "critical point" is precisely this agreement! Apart from the two unhelpful example sentences, such nonsensical explanations would take too much time and effort to reformulate and make useful for the specific purpose. We therefore decided to ditch Ernie Bot for the time being and continue with ChatGPT alone.

This disappointing conclusion could be due to a number of reasons, firstly, as mentioned in Section 2, that Ernie Bot still lags behind ChatGPT from a technological or programming point of view. But it could also be that this large language model has not been trained on a sufficient amount of English grammar texts compared to its Western counterpart, and that the result would be different if the purpose was to explain Chinese grammar. To test this last hypothesis, we asked both chatbots to write Chinese explanations of some grammatical issues in Chinese texts, similar to those in English. However, the respective responses show that ChatGPT is also qualitatively a step ahead of Ernie Bot when it comes to explaining Chinese grammar, so the hypothesis turned out to be false, or at least premature.

4.2 ChatGPT passes the test

ChatGPT was tested with its version 4o. Once we learned how to prompt it in the most appropriate way, the chatbot's responses were generally of a quality that could be easily used for our purposes, although varying degrees of editing were required. For instance, in some cases, both in English and Chinese, the explanations generated contained some "noise" with comments and even whole sen-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1944 (Article)

tences that only served to obscure the message without adding anything new and important to the specific issue being addressed:

— This ensures that the sentence is grammatically accurate and clearly conveys the intended meaning.
— This agreement ensures clarity and accuracy in the sentence.
— This ensures that the subject and verb agree, making the sentence clearer and grammatically correct.
— This can lead to confusion and grammatical errors.
— This common mistake can lead to grammatical errors.

The above are just a few examples of excessive verbosity that does not fit the genre and purpose. The last sentence is even nonsensical, since a mistake can not "lead to" an error, but is by definition an error. However, such unnecessary verbiage, which makes the explanations too long and therefore less readable and didactic, could easily be deleted in a few seconds.

In other cases, the example sentences also contained unnecessary wording that could distract learners from the real issue. An example of this is when ChatGPT was asked to explain the incorrect use of a singular verb with a plural noun as subject:

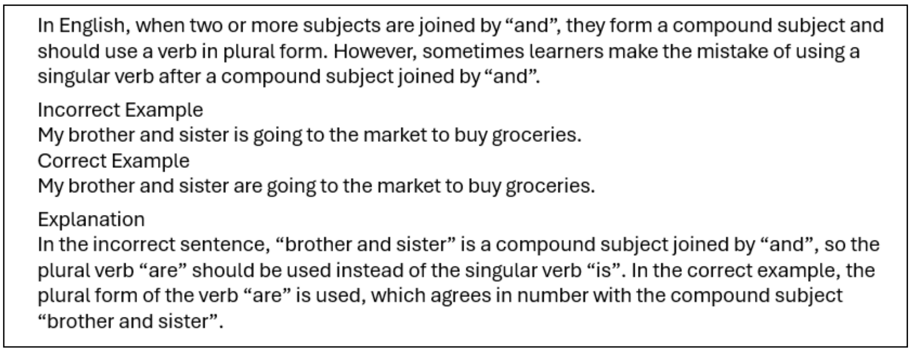
— The dogs runs in the park every morning and play with each other happily.

This sentence includes two verbs (*runs* and *play*) that refer to the same subject (*dogs*), but only the first verb has the incorrect singular form, which requires explanation. The second part of the sentence is therefore superfluous and likely to confuse learners. Therefore, to present a more didactic example, this part of the sentence was removed leaving only the following short incorrect sentence to be corrected:

— The dogs runs in the park every morning.

Finally, ChatGPT was sometimes sloppy in its formulations, for example when it correctly stated in various explanations that a verb should match its subject, but did not make it clear that this match was in terms of number. Again, this problem could easily be solved by simply adding "in number" to avoid any misunderstanding. And a similar method could be used to correct other minor but distracting inaccuracies in the wording.

All in all, despite its shortcomings, ChatGPT has passed the test, thanks to its fruitful symbiosis with experienced and open-minded lexicographers who patiently instruct and correct it, always having the last word. After the initial, somewhat frustrating process of learning how to handle it, the time needed to write the desired explanations was significantly reduced, and in most cases considered to be much shorter than if these explanations had to be conceived from scratch without the inspiration of this generative AI tool. However, this does not exclude that in a very small number of cases a partial or complete reworking of the explanations generated by the chatbot could be beneficial in order to improve their quality, but even this appeared to be relatively straightforward, as the chatbot's original explanation was an inspiration in both a positive and negative sense. Figures 1 and 2 show examples of an original ChatGPT-generated and a modified explanation of noun subject–verb agreement, aimed primarily at learners at the beginner level.



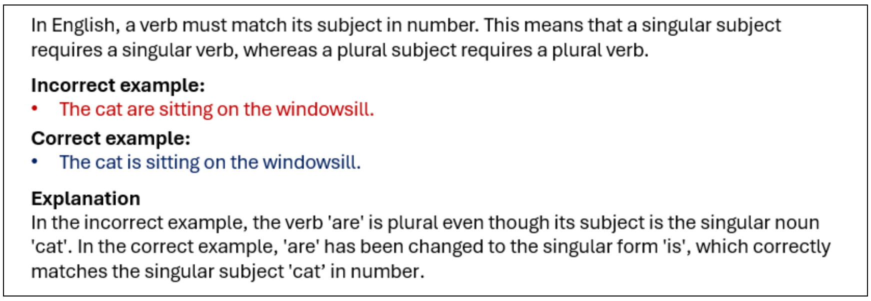
In English, when two or more subjects are joined by "and", they form a compound subject and should use a verb in plural form. However, sometimes learners make the mistake of using a singular verb after a compound subject joined by "and".

Incorrect Example
My brother and sister is going to the market to buy groceries.
Correct Example
My brother and sister are going to the market to buy groceries.

Explanation
In the incorrect sentence, "brother and sister" is a compound subject joined by "and", so the plural verb "are" should be used instead of the singular verb "is". In the correct example, the plural form of the verb "are" is used, which agrees in number with the compound subject "brother and sister".

Figure 1: ChatGPT's explanation of noun subject–verb agreement

Apart from the obligatory layout improvements, there are three major differences between the modified explanation in Figure 2 and the original one in Figure 1. First, the introduction has been completely rewritten and now includes a short presentation of the general rule for subject–verb agreement. Second, the example sentences have been shortened to put more focus on the specific problem, by deleting the superfluous present participial phrase *enjoying the warm sunlight*. And third, the argumentation structure in the final explanation has been adjusted so that the reference to the incorrect example deals only with the error, while the reference to the correct example focuses on how it has been changed to achieve a grammatically correct sentence.



In English, a verb must match its subject in number. This means that a singular subject requires a singular verb, whereas a plural subject requires a plural verb.

Incorrect example:
- The cat are sitting on the windowsill.

Correct example:
- The cat is sitting on the windowsill.

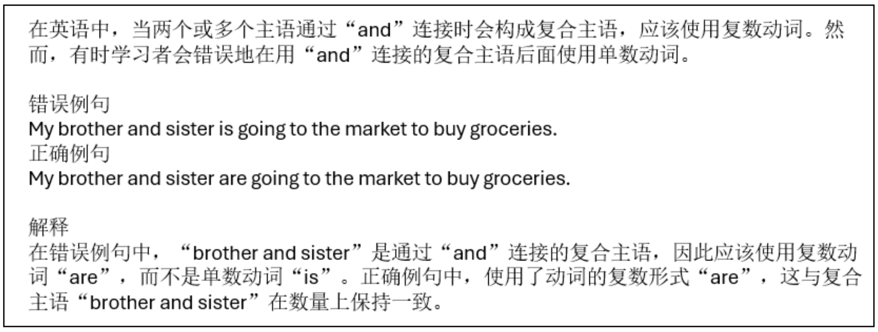
Explanation
In the incorrect example, the verb 'are' is plural even though its subject is the singular noun 'cat'. In the correct example, 'are' has been changed to the singular form 'is', which correctly matches the singular subject 'cat' in number.

Figure 2: Modified explanation of noun subject–verb agreement

4.3 Writing Chinese explanations

After the first round of testing, it was apparent that ChatGPT performed just as well in Chinese as it did in English when asked to explain typical learner errors in English grammar. Since it has probably not been trained on the same amount of Chinese as English texts dealing with English grammar, this is most likely due to some sort of internal translation process, as the explanations in Chinese, and even the example sentences provided, were almost identical to those in English. In any case, once this became clear, we decided to continue with prompts and explanations in Chinese only. Thus, the examples provided in this section were all originally written in Chinese and then translated for the benefit of English-speaking readers.

Figure 3 reproduces ChatGPT's response in Chinese after being prompted to explain why it is a mistake to pair singular verbs with compound subjects, and Figure 4 then provides the translation of this explanation.

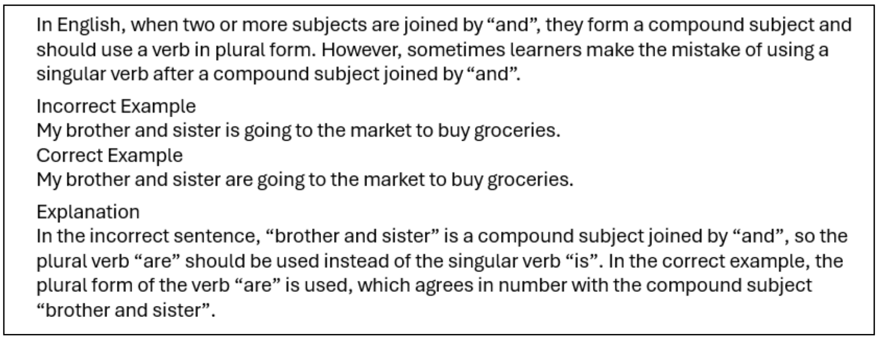


在英语中，当两个或多个主语通过“and”连接时会构成复合主语，应该使用复数动词。然而，有时学习者会错误地在用“and”连接的复合主语后面使用单数动词。

错误例句
My brother and sister is going to the market to buy groceries.
正确例句
My brother and sister are going to the market to buy groceries.

解释
在错误例句中，“brother and sister”是通过“and”连接的复合主语，因此应该使用复数动词“are”，而不是单数动词“is”。正确例句中，使用了动词的复数形式“are”，这与复合主语“brother and sister”在数量上保持一致。

Figure 3: Chinese explanation of compound subject–verb agreement



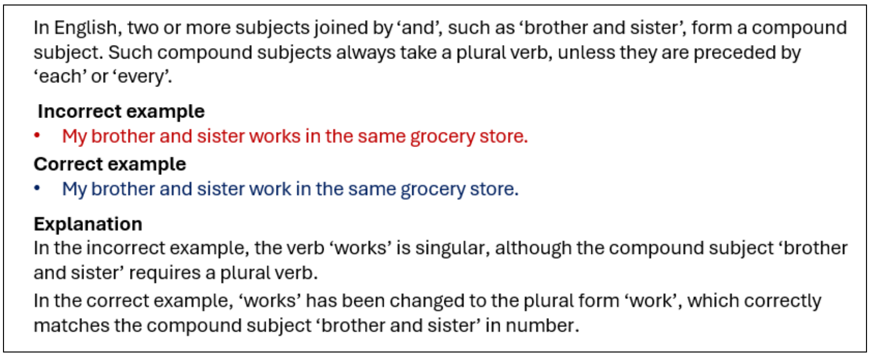
In English, when two or more subjects are joined by “and”, they form a compound subject and should use a verb in plural form. However, sometimes learners make the mistake of using a singular verb after a compound subject joined by “and”.

Incorrect Example
My brother and sister is going to the market to buy groceries.
Correct Example
My brother and sister are going to the market to buy groceries.

Explanation
In the incorrect sentence, “brother and sister” is a compound subject joined by “and”, so the plural verb “are” should be used instead of the singular verb “is”. In the correct example, the plural form of the verb “are” is used, which agrees in number with the compound subject “brother and sister”.

Figure 4: Translation of ChatGPT's explanation in Figure 3

As can be seen in Figure 4, there are some good points in ChatGPT's response, especially the initial explanation of what is meant by *compound subject* and what this type of subject requires of the paired verb in terms of number. However, given its intended audience and genre-specific purpose, there are several things that could and should be improved, as a comparison with the edited explanation in Figure 5 will clearly show. Firstly, there is some distracting "noise" that should be removed, such as the superfluous sentence beginning "However …" in the introduction and the redundant repetition of the definition of a compound subject in the final explanation. Secondly, the compound subject *brother and sister* used in the example sentences has been inserted into the definition of this grammatical category to illustrate what it refers to. Thirdly, the only two words (*each* and *every*) that override the general rule when they precede a compound subject are briefly mentioned. Fourthly, the argumentation structure in the final explanation has been made more logical and straightforward, similar to the refinement of the explanation in Figure 2. As the cherry on the cake, the two example sentences have also been modified, although this was not strictly necessary. The intention behind this move was simply to use a verb other than *to be*, as it turned out to be quite easy to come up with an alternative inspired by the words *groceries* and *market* in the original examples. Figure 5 shows the result of this careful editing, always keeping in mind the specific purpose and anticipated target users.



In English, two or more subjects joined by 'and', such as 'brother and sister', form a compound subject. Such compound subjects always take a plural verb, unless they are preceded by 'each' or 'every'.

Incorrect example

- My brother and sister works in the same grocery store.

Correct example

- My brother and sister work in the same grocery store.

Explanation

In the incorrect example, the verb 'works' is singular, although the compound subject 'brother and sister' requires a plural verb.

In the correct example, 'works' has been changed to the plural form 'work', which correctly matches the compound subject 'brother and sister' in number.

Figure 5: Modified version of ChatGPT's explanation in Figure 3

The revision and editing of the other test explanations followed the same general pattern as the one discussed above in relation to Figures 3, 4 and 5.

- — A quick read through of the AI-generated explanations.
- — Remove some background noise from the text so as not to distract the learner's attention from the main point.
- — Change some words and phrases to improve readability.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1944 (Article)

— Select relevant data, such as subjects and verb forms, from the example sentences and add them to the introduction whenever it makes this section easier to read and helps to clarify the grammar problem being explained, especially if the use of some technical terminology is unavoidable.
— Refine the argumentation structure in the explanation of the incorrect and correct example sentences to make this section more logical, straightforward and concise.
— Improve the layout of the whole explanation to make it as easy as possible for the reader to quickly gain the necessary overview and grasp the essence of the problem being addressed.

As for the example sentences generated by the chatbot, although a few of them were shortened by deleting irrelevant wording to maximise the focus on the specific issue, in only one case was it considered beneficial to rephrase them, as in the explanation in Figure 5. It obviously took some thought, discussion and practice to develop and become familiar with the method described, but once it was internalised the whole editing process became quite straightforward and could be completed in much less time than it would have taken to write the explanations from scratch without inspiration from the chatbot. Although it is for others to judge, the end result can be considered both satisfactory and of the required quality.

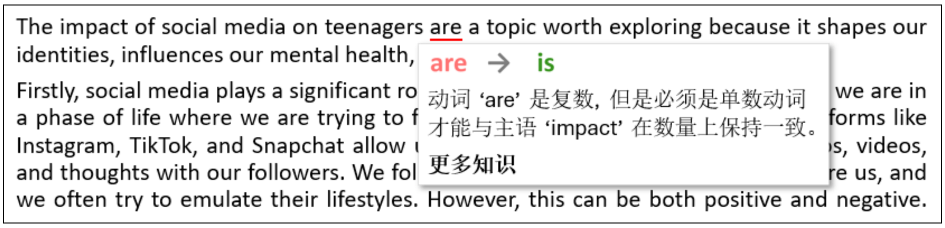
4.4 Explanations at work

The final destination of the explanations discussed in the previous sections is their integration into a bilingual English writing assistant, similar to the bilingual Spanish one presented by Li et al. (2024), but different from theirs in that it is aimed exclusively at Chinese learners of English. The writing assistant will be supported by an AI-powered language model that has been trained to detect errors in written English. The idea is that beginner and intermediate learners can either paste their English texts into the tool or use it to write them, as in Chinese apps of the 1–7 Zuoye type, albeit with different functionality. The writing assistant will then highlight possible problems in the text. If the learner does not know why a particular word has been highlighted, he or she can simply click on the word in question to display a pop-up window with an alternative suggestion followed by a short explanation (see Figure 6). As can be seen, the problem in this case is noun subject–verb disagreement. Accordingly, the short Chinese text reads:

> The verb 'are' is plural, but must be singular to agree in number with the subject 'impact'.

This combination of highlighting, alternative suggestion and ultra-short explanation enhances the possibility of incidental learning as defined from a lexicographic perspective by Tarp (2022).

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1944 (Article)



The impact of social media on teenagers are a topic worth exploring because it shapes our
identities, influences our mental health,
Firstly, social media plays a significant ro … we are in
a phase of life where we are trying to f … forms like
Instagram, TikTok, and Snapchat allow … s, videos,
and thoughts with our followers. We fol … re us, and
we often try to emulate their lifestyles. However, this can be both positive and negative.

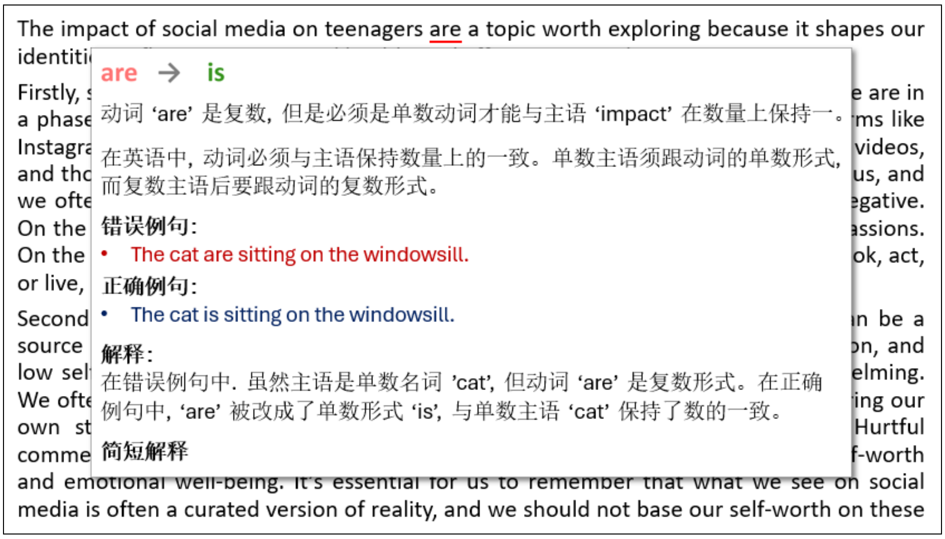
are → is

动词 ‘are’ 是复数，但是必须是单数动词才能与主语 ‘impact’ 在数量上保持一致。

更多知识

Figure 6: Pop-up window with alternative suggestion and short explanation

If the learner is satisfied with the information provided, a simple click on the green *is* will insert that verb form into the text instead of the highlighted *are*. If, on the other hand, the learner is a beginner who wants to know more about this fundamental grammatical issue in English, a click on the Chinese characters 更多知识 (LEARN MORE) in the bottom left corner of the pop-up window will open another window with a long supplementary explanation (see Figure 7).



The impact of social media on teenagers are a topic worth exploring because it shapes our
and emotional well-being. It's essential for us to remember that what we see on social
media is often a curated version of reality, and we should not base our self-worth on these

are → is

动词 ‘are’ 是复数，但是必须是单数动词才能与主语 ‘impact’ 在数量上保持一。

在英语中，动词必须与主语保持数量上的一致。单数主语须跟动词的单数形式，而复数主语后要跟动词的复数形式。

错误例句：

- The cat are sitting on the windowsill.

正确例句：

- The cat is sitting on the windowsill.

解释：
在错误例句中．虽然主语是单数名词 ’cat’，但动词 ‘are’ 是复数形式。在正确例句中，‘are’ 被改成了单数形式 ‘is’，与单数主语 ‘cat’ 保持了数的一致。

简短解释

Figure 7: Pop-up window with supplementary explanation

The first line of the explanation in Figure 7 is a repetition of the short explanation from the pop-up window in Figure 6. The following text is the Chinese version of the English explanation shown in Figure 2, and the characters 简短解释 (LESS) in the bottom left corner indicate how to close the pop-up window.

The whole construction is based on the dialectical relationship between the Hegelian concepts of the *individual*, the *particular* and the *universal*. The under-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1944 (Article)

lined error "are" in the learner's text represents the individual, the short explanation the particular and the long explanation the universal. In this way the particular, i.e. the short explanation, acts as a mediator or bridge between the individual and the universal, allowing the learner to relate the long explanation to the error he or she has made and vice versa. This seemingly simple construction, but with complex underlying relationships, allows for intentional learning of English grammar and is primarily aimed at the motivated student who is eager to study, learn and make progress in English language acquisition.

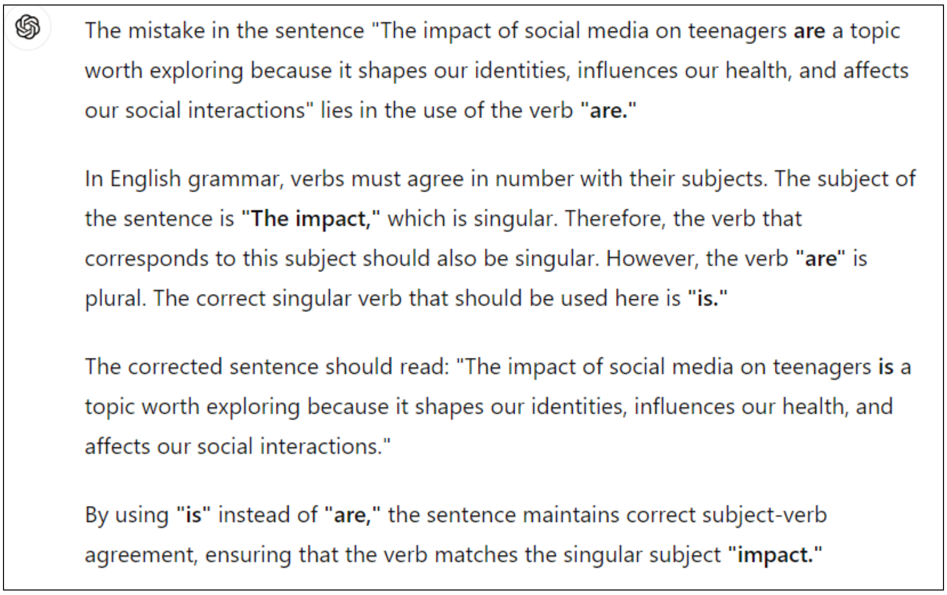
4.5 In the borderland between the possible and the impossible

It should be emphasised that the short explanation in Figure 6 does not use phrases such as *It seems that*, *There may be* and *It looks like*, which have been employed in tools like Grammarly and ProWritingAid, at least until recently, and also by Li et al. (2024) in their proposal for a future writing assistant. The reason for using these and similar phrases is to avoid misinforming the user, as the underlying language models have not been entirely reliable in identifying possible errors and suggesting alternatives. This is about to change. The short explanation proposed in Figure 6 not only states directly that there is an error, but also explains the nature of the problem in very few words.

This new approach is driven not only by improved language models, but also by the integration of generative AI into the writing tool to support its functionality. To explore these new technological advances in the current borderland between the possible and the impossible, ChatGPT was asked to explain why it is a mistake to use the verb form *are*, highlighted as an error by the language model, in the sentence partially covered in Figures 6 and 7. Its response is shown in Figure 8, and as can be seen, it was perfectly able to identify both the subject (*impact*) and its number (*singular*) and compare it with the number (*plural*) of the verb (*are*), and on this basis explain the grammatical problem at hand.

To check whether this was more than a lone wolf, ChatGPT was then asked to explain errors belonging to the 26 sub-categories of subject–verb disagreement listed in Section 3. In all cases, without exception, it managed to identify both the subject and the verb, as well as their respective numbers, and explain the nature of the problem. Although 26 tests are not statistically sufficient to draw a definitive conclusion, the result (26 hits out of 26 possible) strongly suggests that generative AI has an important role to play in improving the quality of writing assistants.

However, as useful as they are, ChatGPT's responses cannot be uploaded directly into small pop-up windows, such as the one in Figure 6, which are intended to give the writer a very brief explanation of why one word or phrase should replace another. They are simply too long for that purpose, and sometimes they are not clear enough or use terminology that learners cannot be expected to know. Nevertheless, the information they provide about a problem and its nature can be of great value in preparing such short default explanations. The best solution would therefore be to prepare an explanation template for each problem category, into which the relevant words, in this case *are* and *impact*, can be inserted, the first one provided by the language model (a technology that already exists) and the second one by the chatbot. How to extract this last information from the chatbot's responses and insert it into the explanations is a technical question. So, the challenge now lies with the computer experts, but lexicographers should also have their say and be prepared for this brave new world. The discussion above is part of that preparation.



The mistake in the sentence "The impact of social media on teenagers **are** a topic worth exploring because it shapes our identities, influences our health, and affects our social interactions" lies in the use of the verb **"are."**

In English grammar, verbs must agree in number with their subjects. The subject of the sentence is **"The impact,"** which is singular. Therefore, the verb that corresponds to this subject should also be singular. However, the verb **"are"** is plural. The correct singular verb that should be used here is **"is."**

The corrected sentence should read: "The impact of social media on teenagers **is** a topic worth exploring because it shapes our identities, influences our health, and affects our social interactions."

By using **"is"** instead of **"are,"** the sentence maintains correct subject-verb agreement, ensuring that the verb matches the singular subject **"impact."**

Figure 8: ChatGPT explains a subject–verb agreement error

5. Conclusions

The main objective of the research project discussed in this paper was "to explore how and to what extent generative AI can be applied to produce different types of explanations that can be activated in writing assistants for Chinese learners of English", based on the hypothesis that this technology can increase productivity without compromising quality.

The paper answered all three questions. It showed *how* generative AI could be used to perform three interrelated tasks in this perspective: (1) determining which error sub-categories should be explained and producing both (2) long and (3) ultra-short explanations of these sub-categories, developing and testing a methodology for each of them.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1944 (Article)

The paper also showed that the use of this technology could (1) significantly speed up the determination of error categories, but probably not to the same quality as if they were based on the much slower method of tagging and then analysing a learner corpus; (2) produce long explanations of the errors much faster, and at least to the same quality as if they had to be written from scratch by human lexicographers; and (3) contribute to the development of much more informative short explanations, and therefore to a radically different quality than those found in writing assistants to date.

Thus, the paper proved that the hypothesis that this technology could increase productivity without compromising quality was correct for long explanations, but not entirely for the prior detection of the error sub-categories to be explained, where quality is likely to be lower, while it was not possible to compare the productivity of short explanations, as those discussed did not yet exist to our knowledge.

Finally, the paper also reported on a comparative test of two different generative AI chatbots, Baidu's Ernie Bot and OpenAI's ChatGPT, both of which were asked to generate explanations in both Chinese and English. Our initial hypothesis that Ernie Bot would be more efficient at writing Chinese than English, and that ChatGPT would be more efficient at writing English than Chinese, proved to be wrong, or at least premature, as ChatGPT performed significantly better than Ernie Bot, with no qualitative difference between its English and Chinese explanations.

As an added bonus, the research project confirmed Tarp and Nomdedeu-Rull's (2024) conclusion that humans should always have "the last word", as generative AI chatbots are not entirely reliable. Like untamed dogs, they need to be kept on a short leash, despite being man's best friend.

One important thing to bear in mind when working with these tools is that they cannot do anything on their own. Beyond their intrinsic technical limitations, their actual performance depends entirely on their interaction with humans, and in particular on the prompts they receive from the latter. This implies that they can only perform at their best when properly handled by a human, in this case a lexicographer. This raises the question of what is required of the lexicographer in order to interact with and prompt the chatbot in an optimal way. Using generative AI chatbots is something that must be learned. Writing good prompts is not easy. Like any learning process, it takes time and a lot of practice. But apart from learning how to handle the chatbots, i.e. acquiring usage skills, it is also important to have a good grasp of the specific domain of knowledge that is the subject of the interaction. Without knowledge of English grammar and the errors that Chinese learners typically make when writing in English, it would be impossible to interact meaningfully with the chatbots and prompt them to improve the explanations they generate. User skills and domain knowledge are the two keys to success in working with this new technology.

Having said that, it is worth remembering the old English saying that dates back to at least the early 17th century: *The proof of the pudding is in the eating*. In this case, the *pudding* is the explanations, the *proof* is testing them, the *eating* is using the writing assistant in which they are integrated, and those who *eat* it are its future target users. If they are not happy and satisfied, the lexicographers and programmers will have to go back to the drawing board. Meeting user needs is the ultimate quality criterion.

Acknowledgments

Special thanks are due to the Center for Lexicographical Studies at Guangdong University of Foreign Studies, for appointing Sven Tarp as Yunshan Chair Professor, thus making the collaboration between the authors of this article possible.

This study is supported by the Research Funding to Qian Li from Center for Linguistics and Applied Linguistics of Guangdong University of Foreign Studies for the project "Multi-modal Language Teaching for Students of Different Disciplines".

References

Abdullayeva, M. and M.Z. Muzaffarovna. 2023. The Impact of Chat GPT on Student's Writing Skills: An Exploration of AI-assisted Writing Tools. *International Conference of Education, Research and Innovation* 1(4): 61-66.

Alonso-Ramos, M. 2023. El papel de ChatGPT como lexicógrafo. Garriga-Escribano, C., S. Iglesia-Martín, J.A. Moreno-Villanueva and A. Nomdedeu-Rull (Eds.). 2023. *Lligams: Textos dedicats a Maria Bargalló Escribà*: 15-27. Tarragona: Publicacions URV.

Alonso-Ramos, M. and M. García-Salido. 2019. Testing the Use of a Collocation Retrieval Tool Without Prior Training by Learners of Spanish. *International Journal of Lexicography* 32(4): 480-497.

Benati, C. and C. Händl (Eds.). 2019. *From Glosses to Dictionaries: The Beginnings of Lexicography*. Newcastle upon Tyne: Cambridge Scholars Publishing.

Bestgen, Y. and S. Granger. 2011. Categorizing Spelling Errors to Assess L2 Writing. *International Journal of Continuing Engineering Education and Life-Long Learning* 21(2/3): 235-252.

Davidson, S., A. Yamada, P. Fernández-Mira, A. Carando, C.H. Sánchez-Gutiérrez and K. Sagae. 2020. Developing NLP Tools with a New Corpus of Learner Spanish. Calzolari, N. et al. (Eds.). 2020. *Proceedings of the Twelfth Language Resources and Evaluation Conference, May 11–16, 2020, Marseille, France*: 7238-7243. Marseille: European Language Resources Association.

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.

Fitria, T.N. 2021. Grammarly as AI-powered English Writing Assistant: Students' Alternative for Writing English. *Metathesis. Journal of English Language, Literature and Teaching* 5(1): 65-78.

Fitria, T.N. 2023. ProWritingAid as AI-Powered Writing Tools: The Performance in Checking Grammar and Spelling of Students' Writing. *Polingua. Scientific Journal of Linguistics, Literature and Language Education* 12(2): 65-75.

Frankenberg-García, A., R. Lew, J.C. Roberts, G.P. Rees and N. Sharma. 2019. Developing a Writing Assistant to Help EAP Writers with Collocations in Real Time. *ReCALL* 31(1): 23-39.

Graham, S. 2020. The Sciences of Reading and Writing Must Become More Fully Integrated. *Reading Research Quarterly* 55(S1): 535-544.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1944 (Article)

Granger, S. and M. Paquot. 2022. *The Louvain English for Academic Purposes Dictionary. User Manual*. Louvain: Centre for English Corpus Linguistics, Université Catholique de Louvain.

Grefenstette, G. 1998. The Future of Linguistics and Lexicographers: Will There Be Lexicographers in the Year 3000? Fontenelle, T., P. Hiligsmann, A. Michiels, A. Moulin and S. Theissen (Eds.). 1998. *Proceedings of the Eighth EURALEX in Liège, Belgium*: 25-41. Liège: English and Dutch Departments, University of Liége.

Gui, S.C. and H.Z. Yang. 2003. *Chinese Learner English Corpus*. Shanghai: Foreign Language Education Press.

Hanks, P. 2012. The Corpus Revolution in Lexicography. *International Journal of Lexicography* 25(4): 398-436.

Hanks, P. 2013. Lexicography from Earliest Times to the Present. Allan, K. (Ed.). 2013. *The Oxford Handbook of the History of Linguistics*: 503-536. Oxford: Oxford University 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.

Hulstijn, J.H. 2013. Incidental Learning in Second Language Acquisition. Chapelle, C.A. (Ed.). 2013. *The Encyclopedia of Applied Linguistics:* 2632-2637. New York: Wiley-Blackwell.

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. *Electronic Lexicography in the 21st Century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 Conference, Brno, 27–29 June 2023*: 518-533. Brno: Lexical Computing CZ s.r.o.

Kilgarriff, A. 2013. Using Corpora as Data Sources for Dictionaries. Jackson, H. (Ed.). 2013. *The Bloomsbury Companion to Lexicography*: 77-96. London: Bloomsbury.

Krashen, S. 1989. We Acquire Vocabulary and Spelling by Reading: Additional Evidence for the Input Hypothesis. *Modern Language Journal* 73: 440-464.

Leow, R.P. and C.C. Zamora. 2017. Intentional and Incidental L2 Learning. Loewen, S. and M. Sato (Eds.). 2017. *The Routledge Handbook of Instructed Second Language Acquisition*: 33-49. New York: Routledge.

Lew, R. 2023. ChatGPT as a COBUILD Lexicographer. *Humanities and Social Sciences Communications* 10:704.

Li, Q., S. Tarp and A. Nomdedeu-Rull. 2024. The Necessary Symbiosis: How ChatGPT Co-authored a New Type of Learner's Grammar. *Círculo de lingüística aplicada a la comunicación* 100. (To appear)

McArthur, T. 1986. *Worlds of Reference. Lexicography, Learning and Language from the Clay Tablet to the Computer.* Cambridge: Cambridge University Press.

McKean, E. and W. Fitzgerald. 2024. The ROI of AI in lexicography. *Lexicography* 11(1): 7-27.

Nomdedeu-Rull, A. and S. Tarp. 2024. *Introducción a la Lexicografía en Español: Funciones y Aplicaciones.* London: Routledge.

Paquot, M. 2012. The LEAD Dictionary-cum-writing Aid: An Integrated Dictionary and Corpus Tool. Granger, S. and M. Paquot (Eds.). 2012. *Electronic Lexicography*: 163-185. Oxford: Oxford University Press.

Phoodai, C. and R. Rikk. 2023. Exploring the Capabilities of ChatGPT for Lexicographical Purposes: A Comparison with Oxford Advanced Learner's Dictionary within the Microstructural Framework. Medved', M., M. Měchura, C. Tiberius, I. Kosem, J. Kallas, M. Jakubíček and S. Krek (Eds.). 2023. *Electronic Lexicography in the 21st Century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 Conference Brno, 27–29 June 2023*: 345-375. Brno: Lexical Computing CZ s.r.o.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1944 (Article)

Rees, G.P. and R. Lew. 2023. 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. 2012. The Road to Automated Lexicography: An Editor's Viewpoint. Granger, S. and M. Paquot (Eds.). 2012. *Electronic Lexicography:* 15-30. Oxford: Oxford University Press.

Rundell, M. 2023. Automating the Creation of Dictionaries: Are we Nearly There? *Asialex 2023 Proceedings. Lexicography, Artificial Intelligence, and Dictionary Users, 22–24 June 2023, Seoul, Korea:* 9-17. Seoul: Yonsei University.

Shu, H., R.C. Anderson and H. Zhang. 1995. Incidental Learning of Word Meanings While Reading: A Chinese and American Cross-cultural Study. *Reading Research Quarterly* 30(1): 76-95.

Song, C. and Y. Song. 2023. Enhancing Academic Writing Skills and Motivation: Assessing the Efficacy of ChatGPT in AI-assisted Language Learning for EFL Students. *Frontiers in Psychology* 14: 1-14.

Tarp, S. 2022. A Lexicographical Perspective to Intentional and Incidental Learning: Approaching an Old Question from a New Angle. *Lexikos* 32(2): 203-222.

Tarp, S. and R.H. Gouws. 2023. A Necessary Redefinition of Lexicography in the Digital Age: Glossography, Dictionography and the Implications for the Future. *Lexikos* 33(1): 425-447.

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* 97: 309-321.

Trap-Jensen, L., H. Lorentzen and N.H. Sørensen. 2014. An Odd Couple — Corpus Frequency and Look-up Frequency: What Relationship? *Slovenščina 2.0: Empirical, Applied and Interdisciplinary Research* 2(2): 94-113.

Wanner, L., S. Verlinde and M. Alonso-Ramos. 2013. Writing Assistants and Automatic Lexical Error Correction: Word Combinatorics. Kosem, I., J. Kallas, P. Gantar, S. Krek, M. Langemets and M. Tuulik (Eds.). 2013. *Electronic Lexicography in the 21st Century: Thinking Outside the Paper. Proceedings of the eLex 2013 Conference, 17–19 October 2013, Tallinn, Estonia*: 472-487. Ljubljana/Tallinn: Institute for Applied Slovene Studies/ Eesti Keele Instituut.

Wiegand, H.E. 1987. Zur handlungstheoretischen Grundlegung der Wörterbenutzungsforschung. *Lexicographica* 3: 178-227.

Wu, L. 2024. AI-based Writing Tools: Empowering Students to Achieve Writing Success. *Advances in Educational Technology and Psychology* 8(2): 40-44.

Xie, Z., G. Genthial, S. Xie, A. Ng and D. Jurafsky. 2018. Noising and Denoising Natural Language: Diverse Backtranslation for Grammar Correction. Walker, M., J. Heng and A. Stent (Eds.). 2018. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL–HLT 2018, New Orleans, Louisiana, USA, June 1–6, 2018, Vol. 1 (Long Papers)*: 619-628. New Orleans: Association for Computational Linguistics.

Zhao, W., L. Wang, K. Shen, R. Jia and J. Liu. 2019. Improving Grammatical Error Correction via Pre-Training a Copy-Augmented Architecture with Unlabeled Data. Burstein, J., C. Doran and T. Solorio (Eds.). 2019. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL–HLT 2019, Minneapolis, MN, USA, June 2–7, 2019. Vol. 1 (Long and Short Papers):* 156-165. Minneapolis: Association for Computational Linguistics.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1945 (Article)

Exploring Dictionary Preferences: A Comparative Study of EFL and GFL Learners in Hungarian Higher Education

Balázs Fajt, *Faculty of Finance and Accountancy,*
Budapest Business University, Hungary
(fajt.balazs@uni-bge.hu) (https://orcid.org/0000-0003-4983-3962)

Abstract: This paper examines the usage patterns of monolingual and bilingual dictionaries among English as a Foreign Language (EFL) and German as a Foreign Language (GFL) learners within the Hungarian higher education context. Despite the prevalent communicative approach in language teaching that often discourages dictionary use in favour of context-based learning, dictionaries are still important resources in acquiring and understanding linguistic nuances and terminology, especially in academic and business settings. Utilizing the quantitative research paradigm, this study gathered data from 371 university students, focusing on their preferences for specific dictionaries and exploring the relationship between their willingness to use these dictionaries and actual usage behaviours. The findings reveal that while EFL learners prefer renowned English monolingual dictionaries, such as Oxford and Cambridge, their willingness to use dictionaries does not necessarily correlate strongly with the frequency of use, suggesting other motivational or contextual influences at play. On the other hand, GFL learners displayed a lower overall engagement with both monolingual and bilingual dictionaries, underscoring potential differences in educational strategies or lower reliance on dictionaries.

Keywords: EFL AND GFL LEARNERS, DICTIONARY PREFERENCES, DICTIONARY USE HABITS, HIGHER EDUCATION, LANGUAGE PEDAGOGY

Zusammenfassung: Erforschung von Wörterbuchpräferenzen: Eine vergleichende Studie von Englisch als Fremdsprache- und Deutsch als Fremdsprache-Lernenden in der ungarischen Hochschulbildung. Dieser Beitrag untersucht die Nutzungsmuster von ein- und zweisprachigen Wörterbüchern unter Lernenden von Englisch als Fremdsprache und Deutsch als Fremdsprache im ungarischen Hochschulkontext. Trotz des vorherrschenden kommunikativen Ansatzes im Sprachunterricht, der die Verwendung von Wörterbüchern oft zugunsten des kontextbezogenen Lernens ablehnt, sind Wörterbücher nach wie vor wichtige Ressourcen für den Erwerb und das Verständnis von sprachlichen Nuancen und Terminologie, insbesondere im akademischen und geschäftlichen Umfeld. Unter Verwendung eines quantitativen Forschungsparadigmas wurden in dieser Arbeit Daten von 371 Universitätsstudenten gesammelt, wobei der Schwerpunkt auf ihren Vorlieben für bestimmte Wörterbücher lag und die Beziehung zwischen ihrer Bereitschaft, diese Wörterbücher zu benutzen, und ihrem tatsächlichen Nutzungsverhalten untersucht wurde. Die Ergebnisse zeigen, dass Englisch als Fremdsprache-Lernende zwar renommierte einsprachige englische Wörterbücher wie Oxford und Cambridge bevorzugen, ihre Bereitschaft zur Nutzung von Wörterbüchern jedoch nicht unbedingt stark mit der Nutzungshäufigkeit korreliert, was auf andere motivationale oder kontextbezogene Einflüsse schließen lässt. Andererseits zeigten Deutsch als Fremdsprache-Lernende insgesamt eine geringere Bereitschaft, sowohl einsprachige als auch zweisprachige Wörterbücher zu benutzen, was auf mögliche Unterschiede in den Lernstrategien oder eine geringere Abhängigkeit von Wörterbüchern hindeutet.

Schlüsselwörter: ENGLISCH ALS FREMDSPRACHE- UND DEUTSCH ALS FREMDSPRACHE-LERNENDE, WÖRTERBUCHPRÄFERENZEN, GEWOHNHEITEN BEIM WÖRTERBUCHGEBRAUCH, HOCHSCHULBILDUNG, SPRACHPÄDAGOGIK

1. Introduction

Language learning is a dynamic and continuous journey, often characterized by diverse methodologies and pedagogical strategies aimed at maximizing learner engagement and proficiency. Among these methodologies, the communicative approach has been widely adopted since the '90s due to its focus on interaction and comprehension within context. However, this approach frequently discourages the use of dictionaries, urging learners to infer the meanings of new words from the context (Adamska-Sałaciak and Kernerman 2016; Augustyn 2013). While contextual learning is undeniably valuable, this discouragement from dictionary use overlooks the benefits that dictionaries can provide as tools for widening linguistic knowledge and enhancing vocabulary acquisition. In addition, the importance of dictionaries extends beyond simple translation (Fuertes-Olivera 2013; P. Márkus 2023); they are crucial for comprehensive language understanding, offering detailed explanations, usage examples, and phonetic information that contextual clues alone may not provide. Furthermore, dictionary skills are a fundamental component of life-long learning in language education (Leaney 2007; P. Márkus 2023), aiding learners not only during formal education but throughout their lives as they encounter new words and expressions.

This article aims to investigate the extent to which English as a Foreign Language (EFL) and German as a Foreign Language (GFL) learners use monolingual and bilingual dictionaries and which specific dictionaries are preferred by learners. Additionally, it seeks to explore the relationship between learners' willingness to use dictionaries and their actual use of popular dictionaries in Hungary. Through this analysis, the study intends to uncover which of these tools could be integrated into language learning (and teaching).

2. Background

In the context under examination in this article, dictionary usage assumes a pivotal role, particularly within business settings. This study specifically focuses on university students from a university in Hungary offering business pro-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1945 (Article)

grammes, where mastering business terminology in both the native language and foreign languages is crucial. For these learners, dictionaries are indispensable tools that facilitate the appropriate understanding and usage of specialized terminology critical to their fields of study. In Hungary, the importance of business English is underscored by the presence of numerous international companies operating within the country. Additionally, due to historical connections and the presence of German companies in the region, the German language is a regionally important foreign language (Csizér and Lukács 2010), thus German proficiency is also highly valued in Central and Eastern Europe. This dual demand shapes the language education landscape, influencing which languages are taught and the resources provided to learners. Furthermore, there is a noteworthy aspect of digital dictionary use to consider — accessibility. Lew (2016) points out that in many parts of the world, there is a low willingness to pay for subscription-based services, let alone paper-based dictionaries. This is also true for the Hungarian context (P. Márkus et al. 2023). Therefore, it is relatively safe to assume that regular users are likely to favour free digital dictionaries. This preference not only reflects broader trends in digital resource usage but also highlights economic considerations that can influence educational tools and their adoption.

3. Monolingual and bilingual dictionaries in L2 learning

In second language acquisition, dictionaries serve as crucial tools, providing learners with essential linguistic resources (Lew 2016; Nesi 2014; Nied Curcio 2022). The choice between monolingual and bilingual dictionaries significantly influences the second language learning process, each type bearing its unique advantages and drawbacks. The subsequent sections elaborate on these advantages and drawbacks, alongside an overview of related previous research in the Hungarian context.

Monolingual dictionaries define words and phrases using exclusively the target language. Their main benefit is that they offer a wide range of lexical and grammatical information (e.g., collocations, countability, etc.) about a given entry This kind of information helps develop a more nuanced understanding of word meanings, usage, connotations, etc., which are often lost in translation. Monolingual dictionaries may also encourage learners to think in the given target language and can also make them realize that meaning very often cannot be expressed in a single word (Thompson 1987). On the other hand, however, they can be challenging for beginners who may not have sufficient language proficiency to fully comprehend definitions (Lew and Adamska-Sałaciak 2015). In addition, the process of understanding definitions in a second language can be time-consuming and may frustrate learners who need quick translations.

Bilingual dictionaries provide translations between the given second language and the learner's first language. These dictionaries are often favoured by beginners and even intermediate learners for their straightforward approach to understanding and vocabulary building (Laufer and Levitzky-Aviad 2006). Bilingual dictionaries offer immediate comprehension of unfamiliar words, making them accessible and user-friendly for all proficiency levels (Nied Curcio 2022). Bilingual dictionaries can accelerate learning by facilitating quicker word recognition and comprehension, which is especially beneficial during early stages of language study (Loucky 2002, 2005). On the other hand, however, there is a risk of developing too much dependency on the learner's first language, which can hinder immersion in the second given language and slow down the acquisition of the language learning process (Baxter 1980). In addition, translations may not always capture the full meaning or cultural nuances of words, potentially leading to misunderstandings or incomplete learning (Thompson 1987). Previous research also shows that learners often prefer bilingual dictionaries over monolingual ones (Atkins and Varantola 1997; Lew 2004; Nesi 2013).

Both dictionary types have their advantages and drawback, but the choice between monolingual and bilingual dictionaries in second language learning should be informed by the learners' proficiency level, learning objectives, and the specific linguistic nuances of the language being studied. Both types of dictionaries have their place in language education, each contributing uniquely to the linguistic and cognitive development of the learner.

4. Previous research on dictionary users

Over the past decade, the number of studies focussing on dictionary users has been gradually increasing. In the international context, studies investigated language teachers, translators and language learners, too (cf. Knežević et al. 2021; Kosem et al. 2019; Müller-Spitzer et al. 2012; Müller-Spitzer 2014; Wolfer et al. 2018). In addition to these studies, further research in recent years (cf. Simonsen 2011; Hult 2012; Lew 2015; Lorentzen and Theilgaard 2012; Müller-Spitzer et al. 2015), has focused on online dictionary usage. In Hungary, less attention has been paid to the use of online dictionaries. Dringó-Horváth (2017) conducted a questionnaire-based study involving 80 university students majoring in German. The largest Hungarian study on online dictionary use was conducted by Gaál in 2016 and 2017, which involved translators and language teachers (Gaál 2016, 2017). In addition, recently another piece of research carried about by P. Márkus et al. (2023) also investigated EFL and GFL majors.

However, it is important to underscore that Varantola (2002: 33) identifies three distinct groups of dictionary users: professional users (e.g., L2 teachers, translators, etc.), non-professional users (e.g., someone simply looking up a word in a dictionary), and language learners. Research predominantly focuses on professional users, a trend that holds true in the Hungarian context, too, with little attention given to language learners. Given this issue, there is a compelling argument for broadening research horizons in order to investigate "general users" as well. This is particularly important because, as Lew (2015) along with Gaál (2020) point out, professional users often engage with languages and dictionaries in a much more sophisticated manner than regular users due to their

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1945 (Article)

academic backgrounds and extensive knowledge about dictionaries. Consequently, conclusions drawn from studies focusing solely on professional users may not be applicable to the average dictionary user, who undeniably significantly outnumber professional users.

5. Research methods

In line with the above theoretical consideration, the following research questions (RQs) were formulated:

- RQ1: What monolingual and bilingual dictionaries do participants prefer to use?
- RQ2: What is the relationship between university students' willingness to use dictionaries and frequency of monolingual and bilingual dictionary use?

In order to find answers to the above research questions, the quantitative research paradigm was adopted and quantitative data were collected using a self-constructed questionnaire relying on previous research (Dringó-Horváth et al. 2020).

5.1 Participants

This study involved a total of 371 participants, who were recruited through purposive sampling from a Hungarian higher education institution. This purposive sampling strategy was aimed at ensuring a relatively homogeneous sample in order to obtain insights into the L2 learning experiences typical of one academic context. Of all participants, 40.7% were male (n=151) and 59.3% were female (n=220). The average age of the participants was 20.59 years, with a standard deviation (SD) of 1.59 years. Participants reported an average of 9.23 years of learning a second language (L2), with a standard deviation (SD) of 4.29 years. Regarding the learned L2s, around half of the participants learn English (49.1%, n=182) and the other half German (51.9%, n=189).

5.2 Research instrument

To assess the willingness of participants to use dictionaries, a multi-item Likert scale consisting of four items was adopted from a previous research paper (Fajt et al. 2024). Participants rated their agreement with each statement on a 5-point scale ranging from 1 (strongly disagree) to 5 (strongly agree). The internal consistency of the items was evaluated using Cronbach's alpha, yielding a coefficient of .750, indicating an acceptable level of reliability. This suggests that the items on the scale are adequately correlated and collectively provide a consistent measure of the construct of willingness to use dictionaries.

For the measurement of the frequency of use of different dictionaries, single-item scales were employed. Single-item scales have been subject to methodological criticism primarily due to concerns about their reliability and validity compared to multi-item scales. Critics often argue that single-item scales may not adequately capture complex constructs because they cannot account for various facets of the construct the way multi-item scales can. However, when a construct is sufficiently narrow and the questions are concrete, such as behavioural questions (i.e., how frequently someone uses a certain type of dictionary) that all respondents understand uniformly, single item scales may be used (Rossiter 2011).

These single item scales targeted various monolingual and bilingual dictionaries that are — based on the results of previous research — prevalent in Hungary. The selection of these dictionaries is justified by their widespread use in the Hungarian context, as identified in a study carried out by P. Márkus et al. (2023). This research paper pinpointed these dictionaries as the most popular ones in a pilot study conducted prior to that study. Each dictionary type was assessed by a specific item asking participants to report the frequency of their dictionary use on a scale from 1 (never) to 5 (very frequently). The dictionaries included both monolingual and bilingual digital dictionaries. English monolingual dictionaries:

1. **Oxford Learner's Dictionaries:** a widely respected source for EFL learners;
2. **Cambridge Learner's Dictionary:** a widely respected source for EFL learners;
3. **Longman Dictionary of Contemporary English:** well-regarded for its clear explanations and reliable example sentences;
4. **The Free Dictionary:** provides a vast range of linguistic resources, including idioms, thesaurus entries, and encyclopaedia facts;
5. **Macmillan Dictionary**[1]: features detailed definitions with particular emphasis on contemporary language and practical usage examples.
6. **Urban Dictionary:** a community-edited dictionary well-known for its informal and slang language content.

German monolingual dictionaries:

1. **Duden:** a widely respected source for GFL language learners;
2. **Langenscheidt:** a source known for its comprehensive coverage of German vocabulary;
3. **Pons:** a source utilized for its clear definitions and usage examples.

Bilingual dictionaries (all of these dictionaries provide both English–Hungarian and German–Hungarian translations):

1. **SZTAKI:** a widely used online Hungarian–English (and Hungarian–German) dictionary; SZTAKI is known for its extensive database, which includes both general language and technical terms. Its user-friendly interface makes it popular among students and professionals alike.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1945 (Article)

2. **DictZone:** this dictionary stands out for its comparison features, allowing users to see multiple translations for a single entry. Its large vocabulary database, including idiomatic expressions and phrases, is particularly useful for language learners looking for nuanced translations.
3. **Akadémiai Publishers:** this dictionary is recognized for its academic precision and comprehensive coverage of both Hungarian–English and Hungarian–German translations. Often used in educational and professional settings, it provides detailed entries with contextual usage, making it suitable for advanced language learners.
4. **Maxim-dictionaries:** as one of Hungary's leading dictionary publishers, Maxim offers both print and digital versions. Its bilingual dictionaries are appreciated for their rich, authoritative content, including cultural and idiomatic expressions, which cater to a wide audience ranging from beginners to advanced learners.
5. **MorphoLogic:** specializing in Hungarian language software, MorphoLogic develops digital bilingual dictionaries that integrate seamlessly with language processing tools. It emphasizes linguistic accuracy and up-to-date terminology, particularly in technical and specialized fields.
6. **English–Hungarian/German–Hungarian dictionary[2]:** Frequently consulted for quick translations, these dictionaries typically feature straightforward entries, making them highly accessible for users seeking rapid, practical solutions for everyday language needs. They are often less detailed but provide immediate, relatively reliable translations.

5.3 Data collection and data analysis

Data for the study were collected during autumn 2023 utilizing a quantitative survey instrument administered online via Google Forms. Participation in the survey was entirely voluntary and anonymous, ensuring that participants could freely decide to engage without any disclosure of their identity. Additionally, participants were informed that they could interrupt filling in the questionnaire at any time without any consequences, allowing them the flexibility to participate according to their comfort and availability.

The data analysis for the study was conducted using both descriptive and inferential statistical techniques. Initially, descriptive statistics such as mean scores and corresponding standard deviations were calculated to provide a basic understanding of the data distributions and central tendencies of the variables studied. For inferential statistics, the study employed two techniques to examine the relationships and differences within the data: independent samples t-tests were used to compare the means of two independent groups on the same continuous, dependent variable. This analysis helped to identify any statistically significant differences between groups (i.e. EFL and GFL learners) within the study. Additionally, Pearson correlation was utilized to assess the strength of potential linear relationships among variables. This analysis was crucial for under-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1945 (Article)

standing how variables related to each other within the context of the study. All statistical analyses were performed using SPSS 28.0 software and all results were considered statistically significant at $p<.05$.

6. Results

The English monolingual dictionary use, as represented in Table 1, displays a range of mean scores indicating varying levels of usage among the dictionaries listed. The Oxford, Urban, and Cambridge dictionaries have relatively higher mean scores (2.32, 2.29, and 2.13, respectively) compared to other dictionaries, suggesting they are more frequently used among participants. These dictionaries also exhibit higher standard deviations (1.49, 1.46, and 1.40, respectively), indicating a greater variability in their usage rates among our respondents. In contrast, The Free Dictionary, Macmillan, and Longman dictionaries have significantly lower mean scores (1.31, 1.20, and 1.19 respectively) and smaller standard deviations (.84, .60, and .64 respectively).

Table 1: English monolingual dictionary use

Dictionary	M	SD
Oxford (https://www.oxfordlearnersdictionaries.com)	2.32	1.49
Urban Dictionary (https://www.urbandictionary.com)	2.29	1.46
Cambridge (https://dictionary.cambridge.org/dictionary/learner-english/)	2.13	1.40
The Free Dictionary (https://www.thefreedictionary.com)	1.31	.84
Macmillan (https://www.macmillandictionary.com)	1.20	.60
Longman (https://www.ldoceonline.com)	1.19	.64

The usage data for German monolingual dictionaries (Table 2) generally indicate lower usage levels across all listed dictionaries compared to the English ones.

Table 2: German monolingual dictionary use

Dictionary	M	SD
Duden (https://www.duden.de/)	1.62	1.28
Pons (http://de.pons.com/)	1.23	.75
Langenscheidt (https://de.langenscheidt.com/)	1.19	.64

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1945 (Article)

The Duden dictionary has the highest mean usage score (1.62) among the German dictionaries, but this is still lower than the top scores in the English dictionary list. Pons and Langenscheidt have lower mean scores (1.23 and 1.19, respectively) and lower standard deviations (.75 and .64, respectively), indicating less frequent and more consistent usage patterns among the respondents.

As a next step, EFL and GFL learners' use of bilingual dictionaries was compared using independent samples t-tests. The p-values obtained for all dictionaries range from .205 to .867 suggesting that the differences in dictionary usage between EFL and GFL learners are not statistically significant.

Table 3: The comparison of bilingual dictionary use among EFL and GFL learners

	EFL (n=182)		**GFL (n=187)**		**t**	**p**	**d**
Dictionaries	M	SD	M	SD			
English/German–Hungarian dictionary	3.26	1.46	3.11	1.70	-.89	.372	.09
SZTAKI	2.04	1.31	2.20	1.42	1.15	.252	.12
Akadémiai Publishers	1.91	1.27	1.88	1.37	-.17	.867	.02
Maxim	1.77	1.32	1.95	1.45	1.27	.205	.13
DictZone	1.76	1.21	1.88	1.48	.86	.392	.09
Morpho-Logic	1.12	.40	1.14	.55	.33	.739	.03

The minimal differences in mean scores and the trivial effect sizes (d) suggest that both groups of language learners may have similar needs and preferences when it comes to dictionaries. In the case of the above dictionaries, these results could imply that factors such as the design and features of dictionaries (e.g., ease of use, quality of translations, comprehensiveness of entries) are likely to be more influential in determining dictionary use than the specific language being learned; however, this should be investigated further through qualitative methods (e.g., interviews).

As a next step, correlation analyses using Pearson's correlation were used to investigate the potential interrelationships among variables. Correlation is a statistical procedure that describes the extent to which two variables are related and "go together". Correlation is expressed as a correlation coefficient (r) ranging from -1 to 1. A correlation coefficient close to 1 indicates a strong positive relationship, meaning as one variable increases, the other also tends to increase. Conversely, a coefficient close to -1 signifies a strong negative (inverse) relationship, in case of which an increase in one variable results in a decrease in the other. A correlation of zero suggests no linear relationship between the variables. It is important to note, however, that correlation does not imply causation; it merely indicates the presence of a relationship between variables, without attributing cause.

Table 4 offers a correlation matrix that explores the relationships between EFL learners' willingness to use dictionaries and their frequency of using various monolingual (ML) and bilingual (BL) dictionaries. In Table 4, only statistically significant correlations are presented.

Table 4: Significant correlations among EFL learners' willingness to use dictionaries and the frequency of use of different monolingual (ML) and bilingual (BL) dictionaries

Scales	1	2	3	4	5	6	7	8	9	10	11	12	13
1. Willingness to use dictionaries	1												
2. Cambridge (ML)	.22	1											
3. Oxford (ML)		.76	1										
4. Urban Dictionary (ML)	.16		.18	1									
5. Macmillan (ML)		.19	.23	.27	1								
6. The Free Dictionary (ML)		.29	.26		.58	1							
7. Longman (ML)	.23	.28	.26		.66	.54	1						
8. SZTAKI (BL)	.29	.27						1					
9. DictZone (BL)	.28	.56	.38			.26		.32	1				
10. Akadémiai Publishers (BL)		.36	.40	.16	.28	.35	.22	.31	.21	1			
11. Maxim-dictionaries (BL)		.24	.21	.19	.23		.22		.18	.26	1		
12. MorphoLogic (BL)					.40	.23	.29				.23	1	
13. English/German–Hungarian dictionary	.17	.26	.23			.21					.21		1

Regarding monolingual dictionaries, Oxford, despite its prominence, shows no correlation with willingness to use dictionaries, suggesting that factors other than general willingness might drive its use. On the other hand, however, Cambridge shows a positive correlation, indicating that learners who are willing to use dictionaries are somewhat likely to use Cambridge, which — in Hungary as well as other parts of the world — is known for its academic credibility. Urban Dictionary exhibits a negligible correlation, possibly because it is often used for informal language learning or specific queries rather than in-school EFL learning purposes. Finally, The Free Dictionary, Macmillan, and Longman show low to no correlation with willingness to use dictionaries. Regarding bilingual dictionaries, SZTAKI, DictZone and the English–Hungarian dictionary show correlation with willingness to use dictionaries but the strength of correlation is negligible even in the case of these dictionaries. The other bilingual dictionaries also show variable correlations with one another ranging from zero to .32. indicating selective preferences among learners.

As a next step, correlation was employed to create a correlation matrix that explores the relationships between GFL learners' willingness to use dictionaries and their frequency of using various monolingual (ML) and bilingual (BL) dictionaries (Table 5). In Table 5, only statistically significant correlations are presented.

Table 5: Significant correlations among GFL learners' willingness to use dictionaries and the frequency of use of different monolingual (ML) and bilingual (BL) dictionaries

Scales	1	2	3	4	5	6	7	8	9	10
1. Willingness to use dictionaries	1									
2. Duden (ML)		1								
3. Langenscheidt (ML)		.42	1							
3. Pons (ML)		.47	.59	1						
2. SZTAKI (BL)					1					
2. DictZone (BL)			.19		.20	1				
2. Akadémiai Publishers (BL)		.22	.34	.22			1			
2. Maxim-dictionaries (BL)			.20		.27		.32	1		
2. MorphoLogic (BL)		.19	.52	.44	.19	.22	.35	.28	1	
2. English–Hungarian/German–Hungarian (BL)		.15	.18	.20						1

The data indicate that there is no correlation between the GFL learner participants' willingness to use dictionaries and their use of any specific dictionary, whether monolingual or bilingual. This suggests that the willingness to engage with dictionaries does not directly influence which dictionaries are used among these GFL learners; this indicates that other factors are at play in dictionary selection. Regarding monolingual dictionaries, they have inter-dictionary correlations (.42 for Duden and Langenscheidt, .47 and .59 for Langenscheidt and Pons respectively), highlighting a pattern of use among those who favour monolingual options. Pons also correlates with Langenscheidt, which implies a preference chain among monolingual dictionaries with users of one being more likely to use the others, perhaps due to their comprehensive coverage and similar linguistic focus. Regarding bilingual dictionaries, MorphoLogic exhibits strong correlations with other bilingual dictionaries and even some monolingual ones (e.g., .52 with Langenscheidt and .44 with Pons). DictZone and Akadémiai Publishers also show correlations with monolingual dictionaries and among themselves. Finally, Maxim-dictionaries and Hungarian–German also demonstrate correlation with other dictionaries.

7. Discussion and implications

The data suggest that there are distinct usage patterns for English and German monolingual dictionaries, which may be influenced by various linguistic, cultural, and educational contexts.

In terms of linguistic features, English dictionaries such as Oxford and Cambridge appear to be more widely used, possibly due to their extensive vocabulary, detailed definitions, and incorporation of idiomatic and colloquial expressions. Their comprehensive entries and inclusion of contemporary usage may make them more appealing for English learners looking to understand nuances in the language. Urban Dictionary, on the other hand, offers a more user-driven approach by focusing on modern slang and informal terms, which could explain its popularity for quick, informal lookups, especially among younger learners or those engaged with English media.

In contrast, German monolingual dictionaries seem to exhibit lower usage rates overall. The limited range of German dictionaries considered in this study could contribute to this pattern, as learners might not find these dictionaries comprehensive enough to meet their language needs. Additionally, the nature of the German language, with its complex grammatical structures and extensive compound words, might make monolingual dictionaries less user-friendly, particularly for beginner and intermediate learners. This linguistic complexity may lead GFL learners to prefer bilingual dictionaries that provide more straightforward translations.

Cultural factors could also play a role in the observed differences. English dictionaries like Oxford, Cambridge, and even Urban Dictionary benefit from a global cultural presence due to the widespread use of English in media, academia, and international communication. This widespread use may lead to a preference for these resources, as learners become familiar with them through cultural exposure.

On the other hand, German dictionaries may not enjoy the same level of international recognition. The lower usage patterns of German dictionaries in this study could reflect a more localized cultural context, where learners of German in Hungary may not feel the same level of cultural connection or necessity to engage with monolingual German dictionaries. This could also stem from a lesser emphasis on German in global media and fewer online resources dedicated to German language learning compared to English.

Educational practices and resources in Hungary could also significantly influence dictionary usage patterns. English dictionaries such as Oxford and Cambridge are often incorporated into the curriculum through widely used coursebooks in primary and secondary schools. Students are introduced to these dictionaries early on, which may create a familiarity effect and a preference for these resources. Furthermore, English language instruction in Hungary generally emphasizes the importance of developing a broad vocabulary and understanding nuanced language use, which aligns with the capabilities of these monolingual English dictionaries.

In contrast, German language instruction might not emphasize monolingual dictionary use to the same extent. The potentially lower motivation levels among GFL learners, as suggested by previous research (Öveges and Csizér 2018), could result in a reduced inclination to use German monolingual dictionaries for language improvement. Additionally, the smaller number of German monolingual dictionaries investigated in this study might limit the available options for learners, potentially influencing the observed usage patterns.

The study found that English monolingual dictionaries like Oxford, Cambridge, and Urban Dictionary are more frequently used and exhibit broader application among learners, likely due to their linguistic comprehensiveness, cultural recognition, and integration into educational practices. German dictionaries, however, showed lower overall usage and greater variability, indicating a more specialized or niche use among GFL learners. These differences may be attributed to the more complex linguistic nature of the German language, cultural factors that do not promote the same level of international usage as English, and educational contexts that do not emphasize German monolingual dictionaries as much.

Compared to monolingual dictionaries, in line with the results of pervious research (Atkins and Varantola 1997; Lew 2004; Nesi 2014), it was also identified here that both EFL and GFL learners demonstrate a higher preference for bilingual dictionaries over monolingual ones. In addition, the analysis of bilingual dictionary usage among EFL and GFL learners revealed subtle but not statistically significant differences in how these dictionaries are used by the two groups. This trend might imply that the choice and usage of bilingual dictionaries are influenced by potential factors that are common to both groups; these may include the accessibility of dictionaries, the learners' perceived ease of use of these tools, or the general educational practices that do not distinctly favour certain dictionaries over others. The lack of significant differences might also suggest that both EFL and GFL learners see equal value in using bilingual dictionaries as a resource for language learning, regardless of the specific language being studied. Given the minimal differences observed, further research could explore other underlying factors that influence dictionary usage, such as individual learner strategies, specific educational contexts, or the design or layout of the dictionaries themselves. Additionally, qualitative methods could provide deeper insights into the subjective preferences and experiences of learners with these tools, potentially uncovering nuanced explanations that quantitative data alone may not reveal.

8. Conclusion

This article sought to explore the extent and manner in which EFL and GFL learners use monolingual and bilingual dictionaries, and to assess the correlation between their willingness to engage with these resources and their actual usage patterns. The findings of this study offer several insights into the dynamics of dictionary use among university students in Hungary, particularly within the context of business language learning.

Regarding the first research question, the data revealed that EFL learners showed a preference for well-established and renowned English monolingual dictionaries, such as Oxford, Cambridge, and Urban Dictionary, which may be useful resources in both academic and informal settings. GFL learners, on the other hand, showed usage across multiple German dictionaries with Duden and Langenscheidt being the most commonly used. Bilingual dictionaries, such as SZTAKI and DictZone were also frequently used by both groups, indicating a broader prevalence in bilingual language contexts. For the second research question, which examined the relationship between university students' willingness to use dictionaries and the frequency of monolingual and bilingual dictionary use, the findings suggest a nuanced landscape. While one might expect a strong correlation between the willingness to engage with dictionaries and their actual usage, the results indicate that this relationship is not straightforward or uniformly strong across different types of dictionaries and languages as well. Among EFL learners, the analysis indicated that there is a relationship between willingness to use dictionaries and the usage of certain bilingual dictionaries, such as SZTAKI and DictZone, which are both commonly used for English–Hungarian translations. This suggests that EFL learners who are more willing to use dictionaries are somewhat more likely to use bilingual resources, potentially to support their language comprehension and translation needs. However, surprisingly, this willingness did not strongly go together with the use of renowned monolingual dictionaries, such as Oxford and Cambridge. This could suggest that EFL learners might not view these prestigious dictionaries as imme-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1945 (Article)

diately necessary for their learning processes, possibly due to sufficient proficiency or alternative learning resources that are less dictionary-dependent or maybe because of lack of skills in how to use a monolingual dictionary. For GFL learners, the analysis did not show significant correlations between willingness and the use of popular bilingual dictionaries or monolingual dictionaries such as Duden and Langenscheidt. This might indicate that GFL learners either rely less on dictionaries as a learning tool or that their willingness to engage with dictionaries does not necessarily translate into frequent use. This could be attributed to different educational approaches, where perhaps a greater emphasis is placed on contextual and immersive learning strategies rather than dictionary-based learning.

The findings suggest that dictionary use among language learners is influenced by a combination of factors. The fact that willingness did not strongly correlate with the use of certain popular dictionaries suggests that motivations for dictionary use are complex and may be driven by specific learning contexts or tasks rather than a general propensity towards using language resources. For future research, it would be beneficial to explore these motivational and contextual factors in more detail, perhaps through qualitative studies that could provide deeper insights into why learners choose to use or not use dictionaries. Such studies could examine the impact of teaching methodologies on dictionary use, learner attitudes towards different types of dictionaries, and the role of dictionaries in developing language competence over time. Furthermore, as digital resources continue to evolve, ongoing evaluation of how digital dictionaries are integrated into language learning curricula could provide valuable feedback for both educational technology developers and language educators aiming to optimize the tools available to learners.

While this study provides valuable insights into the use of dictionaries among EFL and GFL learners in Hungary, several limitations should be acknowledged. Participants were exclusively recruited from a single Hungarian higher education institution, which may limit the generalizability of the findings. Learners from different educational backgrounds might exhibit different patterns of dictionary use. In addition, while the study included several commonly used dictionaries, it did not encompass all possible dictionary options available to learners, potentially overlooking emerging or less mainstream resources that might also be useful resources. Finally, while useful for establishing broad patterns and correlations, quantitative research does not capture the nuanced reasons behind learners' preferences and behaviours. Qualitative data — as explained previously — could provide deeper insights into the motivations and contextual factors influencing dictionary use.

Endnotes

1. Macmillan Dictionary has not been available online since 30 June 2023.
2. https://angol-magyar-szotar.hu/ and https://nemet-magyar-szotar.hu/

References

Adamska-Sałaciak, A. and I. Kernerman. 2016. Introduction: Towards Better Dictionaries for Learners. *International Journal of Lexicography* 29(3): 271-278.
https://doi.org/10.1093/ijl/ecw033

Atkins, B.T.S. and K. Varantola. 1997. Monitoring Dictionary Use. *International Journal of Lexicography* 10(1) 1-45.
https://doi.org/10.1093/ijl/10.1.1

Augustyn, P. 2013. *No Dictionaries in the Classroom:* Translation Equivalents and Vocabulary Acquisition. *International Journal of Lexicography* 26(3): 362-385.
https://doi.org/10.1093/ijl/ect017

Baxter, J. 1980. The Dictionary and Vocabulary Behavior: A Single Word or a Handful? *TESOL Quarterly* 14(3): 325-336.

Csizér, K. and G. Lukács. 2010. The Comparative Analysis of Motivation, Attitudes and Selves: The Case of English and German in Hungary. *System* 38(1): 1-13.
https://doi.org/10.1016/j.system.2009.12.001

Dringó-Horváth, I. 2017. Digitális szótárak — szótárdidaktika és szótárhasználati szokások [Digital Dictionaries — Dictionary Didactics and Usage Habits]. *Alkalmazott Nyelvtudomány Különszám* 2017: 1-27.
http://dx.doi.org/10.18460/ANY.K.2017.005

Dringó-Horváth, I., K. P. Márkus and B. Fajt. 2020. Szótárhasználati ismeretek vizsgálata német és angol szakot végzettek körében. *Modern Nyelvoktatás* 26(4): 16-38.

Fajt, B., K. P. Márkus and M. Bánhegyi. 2024. Attitudes towards Dictionary Use: A Case Study of University Students Studying Languages for Business. *Journal of Adult Learning, Knowledge and Innovation* 7(1): 38-48.
https://doi.org/10.1556/2059.2024.00101

Fuertes-Olivera, P.A. 2013. The Theory and Practice of Specialised Online Dictionaries for Translation. *Lexicographica* 29: 69-91.
https://doi.org/10.1515/lexi-2013-0006

Gaál, P. 2016. Online-szótár-használat Magyarországon (OHM) — egy kérdőíves szótárhasználati felmérés eredményei I [Online Dictionary Use in Hungary — Results of a Questionnaire-based Dictionary Use Survey I]. *Alkalmazott Nyelvtudomány* 16(2): 1-21.
http://dx.doi.org/10.18460/ANY.2016.2.003

Gaál, P. 2017. Online-szótár-használat Magyarországon (OHM) — egy kérdőíves szótárhasználati felmérés eredményei II [Online Dictionary Use in Hungary — Results of a Questionnaire-based Dictionary Use Survey II]. *Alkalmazott Nyelvtudomány* 17(1): 1-19.
http://dx.doi.org/10.18460/ANY.2017.1.005

Gaál, P. 2020. Középiskolás tanulók szótárhasználati szokásai — Egy vas megyei kérdőíves felmérés eredményei [Dictionary Use Habits of Secondary School Students — Results of a Questionnaire Survey in Vas County]. *Alkalmazott Nyelvtudomány* 20(2): 1-19.
http://dx.doi.org/10.18460/ANY.2020.2.006 or https://doi.org/10.1093/ijl/ecv010

Hult, A.-K. 2012. Old and New User Study Methods Combined — Linking Web Questionnaires with Log Files from the Swedish Lexin Dictionary. Vatvedt Fjeld, R. and J.M. Torjusen (Eds.). 2012. *Proceedings of the 15th EURALEX International Congress 2012, 7–11 August 2012, Oslo:* 922-928. Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1945 (Article)

Knežević, L., S. Halupka-Rešetar, I. Miškeljin and M. Milić. 2021. Millennials as Dictionary Users: A Study of Dictionary Use Habits of Serbian EFL Students. *Sage Open* 11(2): 1-11. https://doi.org/10.1177/21582440211008422

Kosem, I., R. Lew, C. Müller-Spitzer, M. Ribeiro Silveira, S. Wolfer et al. 2019. The Image of the Monolingual Dictionary across Europe: Results of the European Survey of Dictionary Use and Culture. *International Journal of Lexicography* 32(1): 92-114. https://doi.org/10.1093/ijl/ecy022

Laufer, B. and T. Levitzky-Aviad. 2006. Examining the Effectiveness of 'Bilingual Dictionaries Plus' — A Dictionary for Production in a Foreign Language. *International of Lexicography* 19(2): 135-155.

Leaney, C. 2007. *Dictionary Activities*. Cambridge: Cambridge University Press.

Lew, R. 2004. *Which Dictionary for Whom? Receptive Use of Bilingual, Monolingual and Semi-bilingual Dictionaries by Polish Learners of English.* Poznan: Motivex.

Lew, R. 2015. Research into the Use of Online Dictionaries. *International Journal of Lexicography* 28(2): 232-253.

Lew, R. 2016. Dictionaries for Learners of English. *Language Teaching* 49(2): 291-294. https://doi.org/10.1017/S026144481500049X

Lew, R. and A. Adamska-Sałaciak. 2015. A Case for Bilingual Learners' Dictionaries. *ELT Journal* 69(1): 47-57. https://doi.org/10.1093/elt/ccu038

Lorentzen, H. and L. Theilgaard. 2012. Online Dictionaries — How Do Users Find Them and What Do They Do Once They Have? Fjeld, R.V. and J.M. Torjusen (Eds.). 2012. *Proceedings of the 15th EURALEX International Congresss, 7–11 August 2012, Oslo:* 654-660. Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo.

Loucky, J.P. 2002. Improving Access to Target Vocabulary Using Computerized Bilingual Dictionaries. *ReCALL* 14(2): 295-314. https://doi.org/10.1017/S0958344002000721

Loucky, J.P. 2005. Combining the Benefits of Electronic and Online Dictionaries with CALL Web Sites to Produce Effective and Enjoyable Vocabulary and Language Learning Lessons. *Computer Assisted Language Learning* 18(5): 389-416. https://doi.org/10.1080/09588220500442764

Müller-Spitzer, C. (Ed.). 2014. *Using Online Dictionaries.* Lexicographica Series Maior 145. Berlin/Boston: Walter de Gruyter.

Müller-Spitzer, C., A. Koplenig and A. Töpel. 2012. Online Dictionary Use: Key Findings from an Empirical Research Project. Granger, S. and M. Paquot (Eds.). 2012. *Electronic Lexicography*: 425-458. Oxford: Oxford University Press.

Müller-Spitzer, C., S. Wolfer, S. and A. Koplenig. 2015. Observing Online Dictionary Users: Studies Using Wiktionary Log Files. *International Journal of Lexicography* 28(1): 1-26. https://doi.org/10.1093/ijl/ecu029

Nesi, H. 2013. Researching Users and Uses of Dictionary. Jackson, H. (Ed.). 2013. *The Bloomsbury Companion to Lexicography*: 62-74. London/New York: Bloomsbury.

Nesi, H. 2014. Dictionary Use by English Language Learners. *Language Teaching* 47(1): 38-55. https://doi.org/10.1017/S0261444813000402

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. Storjohann (Eds.). 2022. *Dictionaries and Society: Proceedings of the XX EURALEX International Congress, 12–16 July 2022, Mannheim, Germany*: 71-84. Mannheim: IDS-Verlag.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1945 (Article)

Öveges, E. and K. Csizér (Eds.). 2018. *Vizsgálat a köznevelésben folyó idegennyelv-oktatás kereteiről és hatékonyságáról* [Study on the Effectiveness of Foreign Language Teaching in Public Education]. Budapest: Oktatási Hivatal.

P. Márkus, K. 2023. *Teaching Dictionary Skills*. Budapest: Károli Gáspár Református Egyetem, L'Harmattan Kiadó.

P. Márkus, K., B. Fajt and I. Dringó-Horváth. 2023. Dictionary Skills in Teaching English and German as a Foreign Language in Hungary: A Questionnaire Study. *Journal of Lexicography* 36(2): 173-194. https://doi.org/10.1093/ijl/ecad004

Rossiter, J.R. 2011. *Measurement for the Social Sciences. The C-OAR-SE Method and Why It Must Replace Psychometrics*. Berlin: Springer.

Simonsen, H.K. 2011. User Consultation Behaviour in Internet Dictionaries: An Eye-Tracking Study. *HERMES — Journal of Language and Communication in Business* 24(46): 75-101. https://doi.org/10.7146/hjlcb.v24i46.97370

Thompson, G. 1987. Using Bilingual Dictionaries. *ELT Journal* 41(4): 282-286. https://doi.org/10.1093/elt/41.4.282

Varantola, K. 2002. Use and Usability of Dictionaries: Common Sense and Context Sensibility? Corréard, M.-H. (Ed.). 2002. *Lexicography and Natural Language Processing. A Festschrift in Honour of B.T.S. Atkins:* 30-44. Göteborg: Euralex.

Wolfer, S., I. Kosem, R. Lew, C. Müller-Spitzer and M. Ribeiro Silveira. 2018. Web-based Exploration of Results From a Large European Survey on Dictionary Use and Culture: ESDexplorer. *Lexikos* 28(1): 440-447.
https://doi.org/10.5788/28-1-1473

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1947 (Article)

Academic Word Families in Online English Dictionaries

Geraint Paul Rees, *Translation and Language Sciences,*
Pompeu Fabra University, Barcelona, Spain
(geraintpaul.rees@upf.edu) (https://orcid.org/0000-0002-9204-8073)

Abstract: The concept of the word family has been widely employed in research on vocabulary in the teaching and learning of foreign and second languages. The underlying assumption being that once learners know one member of a word family, they can recognise other members. Empirical research supports this vis-à-vis receptive knowledge of inflectionally related wordforms. However, studies of academic writing indicate that using appropriate derivative forms of a known word is challenging, suggesting a need for dictionaries with morphological support for writers. Traditionally, in paper-based dictionaries, this need could not be fulfilled due, in part, to space constraints. This study aims to establish if it is met in five online English dictionary websites. It analyses the treatment of seventy-four academic wordforms which academic writers have been shown to have difficulty deriving when presented with the related base word. Results indicate good coverage of the derivative forms across the dictionary websites examined but inconsistency within and between resources in the way in which forms are treated. Differences include the status as entries or subentries and the provision of writing support features such as examples, grammar patterns, and collocation information. Finally, changes to the treatment of derivatives to better serve academic writers are suggested.

Keywords: ACADEMIC WRITING, DERIVATIVE FORMS, LEXICOGRAPHY, MORPHOLOGY, ONLINE DICTIONARIES, VOCABULARY ACQUISITION, WORD FAMILIES, WRITING SUPPORT

Opsomming: Akademiese woordfamilies in aanlyn Engelse woordeboeke. Die woordfamilie-konsep is reeds wyd in woordeskatnavorsing in die onderrig en aanleer van vreemde en tweede tale ingespan. Die onderliggende aanname word gemaak dat wanneer leerders een lid van 'n woordfamilie ken, hulle ook ander lede kan herken. Empiriese navorsing steun hierdie aanname ten opsigte van reseptiewe kennis van fleksieverwante woordvorme. Studies van akademiese skryfwerk toon egter dat die gebruik van toepaslik afgeleide vorme van 'n bekende woord 'n uitdaging bied, wat daarop dui dat daar 'n behoefte aan woordeboeke met morfologiese steun vir skrywers bestaan. Tradisioneel kon, deels weens ruimtebeperkings, nie aan hierdie behoefte in papiergebaseerde woordeboeke voldoen word nie. In hierdie studie word beoog om vas te stel of daar in vyf aanlyn Engelse woordeboekwebtuistes wel hieraan voldoen word. Die hantering van vier-en-sewentig akademiese woordvorme waarmee akademiese skrywers sukkel om afleidings daarvan te vorm wanneer hulle die verwante basiswoord teëkom, word geanaliseer. Die resultate dui op goeie verteenwoordiging van die afgeleide vorme in die woordeboekwebtuistes wat ondersoek is, maar toon ook teenstrydighede binne en tussen hulpbronne t.o.v. die metode waarop die vorme hanteer word. Verskille sluit die status as inskrywings of subinskrywings en die voorsiening

Lexikos 34 (AFRILEX-reeks/series 34: 2024): 437-468

van skryfhulpkenmerke soos voorbeelde, grammatikale patrone en kollokasie-inligting in. Ten slotte word veranderings aan die hantering van afleidings voorgestel om akademiese skrywers beter van hulp te kan wees.

Sleutelwoorde: AKADEMIESE SKRYFWERK, AFGELEIDE VORME, LEKSIKOGRAFIE, MORFOLOGIE, AANLYN WOORDEBOEKE, WOORDESKATVERWERWING, WOORDFAMILIES, SKRYFHULP

Introduction

Over the last three decades, the term 'word family' has been used in language teaching and vocabulary research to describe the categorisation of wordforms based on their inflectional and derivational morphology. The construct has been adopted enthusiastically in research on vocabulary in English language teaching. A key factor motivating the concept of word family (henceforth WF) was a desire to provide guidelines for the treatment of morphologically related wordforms in lexicography and language teaching (Bauer and Nation 1993). The starting point for this study is a list of seventy-four wordforms frequently used in academic contexts. It comprises sixteen basic wordforms and their derivatives. Empirical research has shown that L2 users have difficulty producing the WF members (i.e., related wordforms) of these forms in writing (Schmitt and Zimmerman 2002). This study aims to establish how well L2 users of the "Big Five" English dictionary websites are supported when producing these problematic forms by examining their treatment on these websites.

Word families and levels. The idea motivating WFs is that wordforms can be grouped based on their inflectional and derivational morphology. These groups can then be organised into levels. Table 1 reproduced from Bauer and Nation (1993: 254) shows the levels for the WFs *develop, wood* and *bright.*

Table 1: Additions to a WF at different levels of inflection and affixation (reproduced from Bauer and Nation (1993: 254))

Word families			
2	develop develops developed developing	wood wood's woods wooded	bright brighter brightest
3	developable undevelopable developer(s) undeveloped	woody woodiest woodier woodiness	brightly brightish brightness

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1947 (Article)

4	development(s) developmental developmentally		
5	developmentwise semideveloped antidevelopment	wooden	brighten
6	redevelopment predevelopment	anti-wood	

An increase in WF level entails greater formal or semantic irregularity. At Level 1, each form represents a distinct word (i.e., one word = one family). At Level 2, inflected forms with the same base are grouped. The idea being that a learner who can recognise and use *develop* or any of its inflected forms could recognise and use the base or any of its other inflected forms. From Levels 3 to 6 eight criteria determine the level of an affix and its derived wordform (Bauer and Nation 1993: 256).

1. Frequency (generalisability): Affixes at lower levels occur in many wordforms. For example, the Level 2 inflectional affixes *-s, -ed, -ing*, are common to all English verbs. In contrast the affixes *pre-* and *re-* are far less generalised.
2. Productivity: The possibility of the affix forming new wordforms. Inflectional affixes *-s, -ed, -ing* frequently produce new forms with the base of any verb. Whereas since *-ful* is far more selective in the nouns and verbs it combines with, it produces far fewer wordforms.
3. Predictability: The extent to which the meaning of the word created by affixation can be predicted from the meaning of the base and the affix. For example, *-ly* attached to an adjective X, typically means 'in X manner', thus is highly predictable. In contrast, *-ful* when attached to nouns does not always produce word with predictable meaning (e.g., *awful weather* ≠ *awe inspiring weather*).
4. Regularity of the written form of the base: At lower levels, removing the affix leaves the base intact, at higher levels orthographic changes to the base are evident (c.f. *red +ness* and *impose + ition*).
5. Regularity of the spoken form of the base: At lower levels removing the affix leaves it phonologically intact, at higher levels phonological accommodations are evident. For example, removal of the Level 6 affix *-ify* from *mystify* gives *myst;* not a free base in its spoken form.
6. Regularity of spelling of the affix (allomorphy): For example, *pre-* has one written form, while *in-*, *im-*, *il-*, and *ir-* are allomorphs of *in-*.
7. Regularity of the spoken form of the affix (allomorphy): The extent to which the phonological form of the affix is predictable. For example, although the Level 1 affix *-ed* has three spoken forms, these are predictable.

8. Regularity of function: The extent to which the affix attaches to a base of a particular word class and produces a word of a particular class. For example, *-ship* always combines with nouns to produce nouns.

By applying the criteria above, Bauer and Nation (1993) produced the list of affixes in Table 2. Two levels are omitted: Level 1 where each wordform is treated as a different WF, and Level 7 where items have classical roots and affixes that are not found in the sample of WFs in this study.

Table 2: Affixes at different WF levels

Level	Affixes
2	*-s, -ed, -ing*
3	*-able, -er, -ish, -less, -ly, -ness, -th, -y, non-, un-*
4	*-al, -ation, -ess, -ful, -ism, -ist, -ity, -ize, -merit, -ous, in-*
5	*-age, -al, -an, -ance, -ant, -ary,-atory, -dom, -eer, -en, -en, -ence, -ent, -ery, -ese, -esque, --ette, -hood, -i, -ian, -ite, -let, -ling, -ly, -most, -ory, -ship, -ward, -ways, -wise, ante-, anti-, arch-, bi-, circum-, counter-, en-, ex-, fore-, hyper-, inter-, mid-, mis-, neo-, post-, pro-, semi-, sub-, un-*
6	*-able, -ee, -ic, -ify, -ion, -ist, -ition, -ive, -th, -y, pre-, re-*

Word families and language teaching and learning. The usefulness of WFs for language teaching relies on the assumption that once learners know one member, they can recognise others. This has been termed relational knowledge (Tyler and Nagy 1989). Some empirical research supports this for L1 readers and inflectionally related wordforms. However, that derived forms are generally acquired after inflected forms suggests they pose greater problems (Berko 1958). For L2 users, the assumption of relational knowledge is more uncertain. Even proficient L2 users find using suitable derived forms of a known word challenging.

Studies on L2 writing or vocabulary acquisition suggest that learners find derivational morphology challenging. A longitudinal study of English vocabulary acquisition involving three L2 English postgraduate students in the UK indicated gaps in participants' morphological repertoire, particularly regarding the formation of adjectives and adverbs. Schmitt (1998) suggests that morphological errors become fossilised since two of the three participants made little progress producing morphologically related forms over an academic year. Another study of the English word association and grammatical suffix knowledge of 95 secondary and undergraduate students of English in Japan found participants gained 330 words on average over an academic year but could only produce 15% of the possible derivatives (Schmitt and Meara 1997). Similarly, in a study

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1947 (Article)

of TOEFL vocabulary involving 30 learners taking English language courses in preparation for undergraduate study in the UK, participants could only produce derivates in all four major word classes for 12 of 180 possible target words (Schmitt 1999).

Research focusing on productive knowledge of derivational morphology among learners is rarer. Schmitt and Zimmerman's (2002) carefully designed study examined the productive knowledge of 106 L2 English students who comprised two groups: One undertaking pre-sessional and undergraduate English language courses at universities in the US and the UK, and another an MA in English language teaching at a university in the UK. Participants were given 16 prompt words for which they were asked to complete gapped sentences by producing derivative forms of the prompt word from the four major word classes (noun, verb, adjective, and adverb). Participants produced only 50% of the derivative forms permissible. Although the presumably more proficient MA group performed better, knowledge of derived word-forms was still partial even for words which participants felt they knew well. This demonstrates a need for dictionaries that support written production of derived wordforms.

Word families and dictionary making. WFs were posited to help lexicographers treat morphology in a principled and consistent way. Bauer and Nation (1993) criticise the inconsistent treatment of derived forms as entries and sub-entries in several general-purpose English dictionaries from the late 1970s and 1980s. They are not alone in highlighting this issue. However, much research has focused on affixes themselves rather than the derivative forms produced by affixation. For example, Stein (1985) highlights different policies on the positioning of affixes in the indexes of several MLDs. Considering dictionaries as writing aids, it makes little sense to focus on affixes themselves rather than derivative wordforms produced by affixation. Writers are unlikely to ask, 'What word can I form with *-ize*?' but will likely query the use of a particular word, for example, 'How do I use *philosophize* in a sentence?'

There is some consensus on the treatment of wordforms derived by affixation. To be included, a derivative form must be established enough to occur above a certain frequency (De Caluwe and Taeldeman 2003; Stein 1985). Semantic predictability is another important consideration: "The more the meaning of a combination is assumed to be inferable from the meaning of its constituents listed in the dictionary and the process of formation itself, the stronger the likelihood that it will not be listed as a dictionary item" (Stein 1985: 38).

Analyses of entries for derivative forms reveal diverse interpretations of these criteria. In an examination of eight monolingual English desk dictionaries including MLDs, Stein (1985) highlights inconsistent definition of *-ish* derivatives from adjectives designating colour, and inconsistent treatment of derived forms as either main entries or run-ons. Similarly, De Caluwe and Taeldeman (2003) demonstrate inconsistent treatment of wordforms derived from *water* in the *Woordenboek der Nederlandsche Taal* (Van Sterkenburg 1992: 115), noting that some

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1947 (Article)

are listed as separate entries or lemmas and others within the headword *water.*

WFs were posited to remedy these inconsistencies. The idea is that as formal and semantic irregularities increase with higher-level word families, they require "more attention" from the lexicographer (Bauer and Nation 1993: 255). Bauer and Nation suggest ignoring regular, semantically transparent wordforms at Level 1; listing those created by inflection affixation at Levels 2 and 3 as non-defined sub-entries and treating higher-level items as main entries.

WFs and electronic lexicography. Electronic lexicography has been suggested as the solution to the inconsistent treatment of derived forms. Firstly, ostensibly freed from space constraints of paper dictionaries, electronic dictionaries have the potential to include information on all the derived wordforms in a language[1]. Secondly, unbound by the alphabetical index, they could offer several routes to the derivative wordform (De Caluwe and Taeldeman 2003). Regarding the first point, De Caluwe and Taeldeman (2003) stress the importance of not overwhelming the reader with information: "it is not the intention to confront the reader with an interminable amount of information, but this should be possible if the reader so desires" (De Caluwe and Taeldeman 2003: 121). Regarding access structure, they sketch an example of how an onomasiological query for "the fact/quality of being long" (De Caluwe and Taeldeman 2003: 123) might proceed in an ideal dictionary. With reference to *Elektronisches Lernerwörterbuch Deutsch–Italienisch/Dizionario Elettronico per Apprendenti Italiano–Tedesco* (ELDIT), Ten Hacken, Abel and Knapp (2006) present a detailed example of how derivative forms can be treated in electronic dictionaries.

Aims. Lexicography has changed significantly since Bauer and Nation's guidelines were published. Many space and alphabetical ordering constraints of paper dictionaries have been mitigated in online resources. These could feasibly accommodate calls from research on WFs in language teaching for greater writing support for L2 English with derived forms. Accordingly, this study aims to investigate how derivatives are represented in online English dictionary websites consulted by learners. It will answer the following research questions:

1. How well are derivationally related members of WFs covered by dictionary websites with online monolingual English dictionaries?
2. To what extent are they treated in a way which facilitates use in writing?

Methodology

In this section, the dictionary websites examined are discussed along with the reasons for their selection. Next, the sample of 74 derived wordforms shown to be problematic for L2 English users is presented and the process Schmitt and Zimmerman (2002) used to obtain this list explained. Finally, the categories and procedure used in this analysis are given.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1947 (Article)

Dictionary websites examined. This study examines the treatment of morphological behaviour on five popular English dictionary websites (Cambridge, https://dictionary.cambridge.org/ (CAM); Collins, https://www.collinsdictionary.com/ (CD); Longman, https://www.ldoceonline.com/ (LONG); Macmillan, https://www.macmillandictionary.com/ (MELD); and Oxford, https://www.oxfordlearnersdictionaries.com/ (OX)[2].). The versions examined were those live in December 2022.

Monolingual Learners Dictionaries (MLDs) are the obvious place to investigate morphological information for learners. However, the migration from paper-based dictionaries to online dictionaries complicates this assumption. Of the "Big Five" monolingual English dictionary makers, only Longman and Macmillan offer direct access to their MLDs. *LDOCEonline.com* also gives access to the *Longman Business Dictionary*[3] (LBD). Access to the MLDs of Cambridge, Collins, and Oxford is offered via portals which aggregate content from several different dictionaries. For example, the *collinsdictionary.com* entry for **precision** collates data from *Collins COBUILD* (COBUILD), *Collins English Dictionary* (CED)[4], *Webster's New World College Dictionary* (Agnes 2010) (WNWCD4), and other ancillary sources. This study investigates the data presented by each portal rather than focusing only on entries from MLDs since, while dictionary researchers are cognisant of different dictionary types and their target users, many end-users, particularly those at lower proficiency levels, simply want to get the job done. It would be strange if an end-user disregarded information from *collinsdictionary.com* because it came from CED not COBUILD.

Productively challenging academic word families. Schmitt and Zimmerman (2002) judge 74 wordforms acceptable responses to gapped sentences based on sixteen prompt words. These represent an ideal sample with which to investigate the treatment of morphological information in online English dictionaries. The sixteen prompt words were selected from Coxhead's (2000) *A New Academic Wordlist* (AWL). This lends content validity since many English dictionary users, including those shown to have problems with derivative forms in the research discussed above, work in academic contexts.

To obtain the list of 74 acceptable derivate wordform responses, Schmitt and Zimmerman (2002) first extracted all listed derivatives from four learners' dictionaries[5]. Secondly, they used frequency information from the BNC1994 to remove infrequent derivatives. Finally, they elicited responses from 36 L1-English university students to the same gapped sentence prompts used by the non-native speakers. In arriving at their list of acceptable responses, they prioritised this final step. Table 3 shows WFs containing the basic and related wordforms along with their word class and WF level in parenthesis.

Table 3: List of problematic academic WFs (adapted from Schmitt and Zimmerman 2002: 168)

Noun	Verb	Adjective	Adverb
assumption (4)	*assume* (2)	assumed (2) X	X
authority (4) authorization (4)	authorize (4)	authorized (2) authoritive (6) authoritative (6)	authoritively (6) authoritatively (3)
tradition (2)	traditionize (4)	*traditional* (4)	traditionally (3)
selection (6)	*select* (2)	selective (6) select (2) selected (2)	selectively (3)
access (2)	*access* (2)	accessible (7) accessed (2)	accessibly (3) X
ethnicity (4)	X	*ethnic* (2)	ethnically (3)
philosophy (2)	philosophize (4)	philosophical (4) philosophic (6)	philosophically (3)
inevitability (4)	X	inevitable (3)	*inevitably* (3)
liberality (3) liberalization (4) liberalness (3) liberty (2)	liberalize (4)	*liberal* (4)	liberally (3)
release (2)	*release* (2)	released (2)	X
survival (4)	*survive* (2)	surviving (2)	X
ideology (2)	X	ideological (4) X	ideologically (3)
precision (6) preciseness (3)	X	*precise* (2)	precisely (3)
minimum (2) minimization (4)	minimize (4)	minimal (4) minimum (2)	minimally (3)
coherence (5) coherency (3)	cohere (2) X	*coherent* (2)	coherently (3)
persistence (5) persistency (3)	*persist* (2)	persistent (5)	persistently (3)

Note. Italics indicate the most frequently occurring member of each WF in BNC1994; X indicates Schmitt and Zimmerman's judgment that no typical form exists.

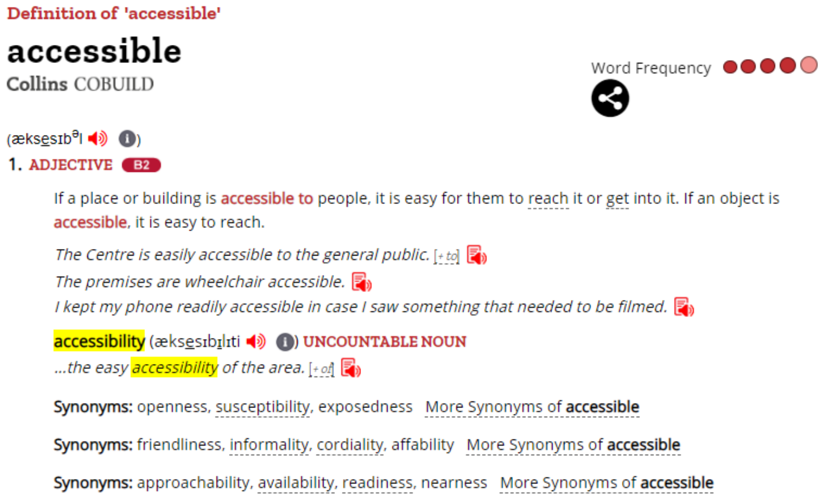
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1947 (Article)

There are, at least, two notable points about this list. Firstly, Schmitt and Zimmerman (2002) treat *accessed*, *assumed, authorized, released* and *surviving* as adjectives. However, the first four could reasonably be verbs and *surviving* could be a verb or a noun. This is a frequent dilemma in English lexical analysis with no satisfactory answer (Hanks 2013). There are cases where these items are used as verbs and others where they are used as adjectives (Frankenberg-Garcia, Rees and Lew 2021). The analysis procedure below accounts for this. Secondly, the AWL has received criticism for ignoring discipline-specific differences in meaning, not accounting for the role of collocates in conditioning meaning and being based on the outdated *A General Service List* (West 1953) (Hyland and Tse 2007; Rees 2021). However, Schmitt and Zimmerman (2002) suggest these are words learners in academic contexts often need to produce. Their standout finding of partial knowledge of derivative forms demonstrates that L2 English users struggle to produce these words. Consequently, these are words for which they could conceivably consult a dictionary for guidance.

Procedure. A search for each of the seventy-four problematic derived forms is conducted on the five dictionary websites. The analysis of the results proceeds in two stages: Stage 1 records whether a wordform is covered; Stage 2 records whether the wordform is treated in a way that supports writing. Namely, whether examples and/or grammar and collocation information are provided. Except for MELD, the websites offer access to several different individual dictionaries. To imitate typical user behaviour, default settings for English language searches are used, and only those dictionaries from which data is presented on the initial results page are considered.

Stage 1: Analysis of coverage. Evaluating dictionary coverage involves not only judging if an item is covered, but also *how* it is covered. This study distinguishes between main entries and sub-entries. Across all the dictionary websites, in main entries the target word is listed as a headword. Sub-entries are more diverse. CAM and OX do not use sub-entries for derived forms. CD often lists derived forms as sub-entries as part of the main entry for the base form (Figure 1). On CD, derived forms are often, simultaneously, presented at the foot of the main entry for the base form under the heading "Derived forms" (Figure 2). In many LONG entries, "Word families" containing derived forms are shown at the top of the results page. On LONG, derived forms are sometimes presented as sub-entries (Figure 3). MELD often lists derived forms at the foot of the main entry for the base word under the heading "Derived word" (Figure 4). For ease of comparison, all these variations in sub-entry presentation are labelled 'sub-entry' here.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1947 (Article)



Definition of 'accessible'

accessible

Collins COBUILD

(æksesɪbᵊl)

1. ADJECTIVE B2

If a place or building is accessible to people, it is easy for them to reach it or get into it. If an object is accessible, it is easy to reach.

The Centre is easily accessible to the general public. [+ to]

The premises are wheelchair accessible.

I kept my phone readily accessible in case I saw something that needed to be filmed.

accessibility (æksesɪbɪlɪti) UNCOUNTABLE NOUN

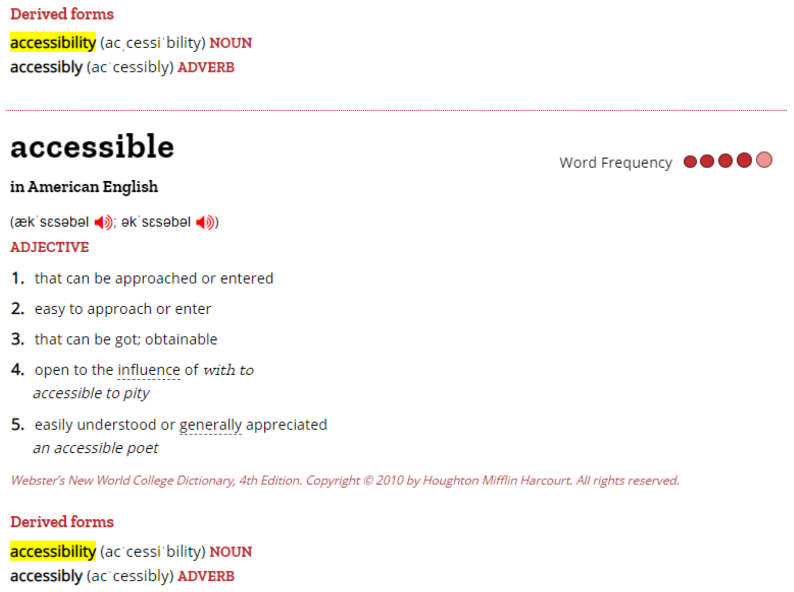
...the easy accessibility of the area. [+ of]

Synonyms: openness, susceptibility, exposedness More Synonyms of **accessible**

Synonyms: friendliness, informality, cordiality, affability More Synonyms of **accessible**

Synonyms: approachability, availability, readiness, nearness More Synonyms of **accessible**

Figure 1: Entry for **accessible** from CD with derived form **accessibility** as sub-entry (highlighted)



Derived forms

accessibility (acˌcessiˈbility) NOUN

accessibly (acˈcessibly) ADVERB

accessible

in American English

(ækˈsɛsəbəl; əkˈsɛsəbəl)

ADJECTIVE

1. that can be approached or entered
2. easy to approach or enter
3. that can be got; obtainable
4. open to the influence of *with to*
   accessible to pity
5. easily understood or generally appreciated
   an accessible poet

Webster's New World College Dictionary, 4th Edition. Copyright © 2010 by Houghton Mifflin Harcourt. All rights reserved.

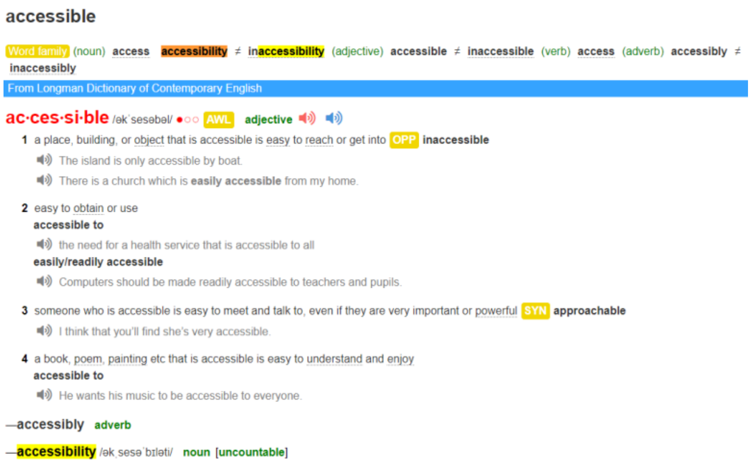
Derived forms

accessibility (acˈcessiˈbility) NOUN

accessibly (acˈcessibly) ADVERB

Figure 2: Entry for **accessible** from CD with derived form **accessibility** as listed under "Derived forms" (highlighted)

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1947 (Article)



accessible

Word family (noun) access accessibility ≠ inaccessibility (adjective) accessible ≠ inaccessible (verb) access (adverb) accessibly ≠ inaccessibly

From Longman Dictionary of Contemporary English

ac·ces·si·ble /əkˈsesəbəl/ ●○○ AWL adjective

1 a place, building, or object that is accessible is easy to reach or get into OPP inaccessible
The island is only accessible by boat.
There is a church which is **easily accessible** from my home.

2 easy to obtain or use
accessible to
the need for a health service that is accessible to all
easily/readily accessible
Computers should be made readily accessible to teachers and pupils.

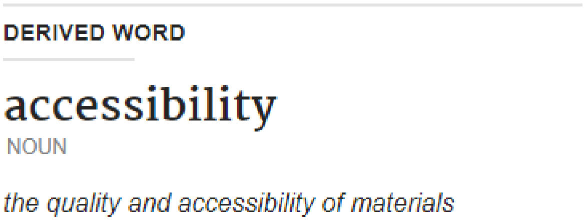
3 someone who is accessible is easy to meet and talk to, even if they are very important or powerful SYN approachable
I think that you'll find she's very accessible.

4 a book, poem, painting etc that is accessible is easy to understand and enjoy
accessible to
He wants his music to be accessible to everyone.

—accessibly adverb

—accessibility /əkˌsesəˈbɪləti/ noun [uncountable]

Figure 3: Entry for **accessible** from LONG with derived form **accessibility** as listed at end of main entry



DERIVED WORD

accessibility

NOUN

the quality and accessibility of materials

Figure 4: Derived form *accessibility* as listed at end of main entry for **accessible** from MELD

Additionally, the websites' response to searches for rare wordforms differs. CAM sometimes uses placeholder examples retrieved automatically from a corpus. If no standard entry can be found, MELD occasionally redirects the user to an example from its crowd-sourced *OPEN DICTIONARY*. In this analysis, placeholder and crowdsourced examples are treated as coverage provided examples are relevant to the target word. CD, LONG, MELD, and OX redirect the user to the more common wordform (e.g., *philosophic* redirects to *philosophical*). The common and rarer forms are considered interchangeable.

The treatment of words with ambiguous word classes, principally *-ed* affixes in the sample, varies within and between the dictionary websites. Searches for wordforms presumed adjectives by Schmitt and Zimmerman (2002) ending in *-ed* often redirect to the entry page for the verb in MELD and OX which use different pages for word classes, and to the main-entry page in covering both noun and verb in the other resources. If a sub-entry exists on these pages for the adjectival form, this form is recorded as being covered (e.g., **authorize** in LBD). Occasionally, traditional examples or corpus lines (automatically generated, occasionally incomplete sentences) illustrate an adjectival use even though the adjectival sense in not explicitly covered (e.g., *release* in CAM: "To what extent the rural sector absorbs the *released labour* is not clear"). In these cases, the presence of a relevant example or corpus line is noted for stage two of the analysis.

Stage 2: Analysis of support for written production. A key assumption of this study is that examples and information about typical grammatical and collocational behaviour support productive use of the wordforms. While there is much research about what constitutes a good dictionary example (Kilgarriff et al. 2008) and the optimal number of examples for supporting production (Frankenberg-Garcia 2015; Ptasznik 2023), here analysis is limited to noting the presence or absence of examples.

In this study, typical combinations containing grammatical words (i.e., prepositions and determiners) are labelled grammar patterns while typical combinations of lexical words are labelled collocations. This policy is maintained irrespective of how these combinations are labelled on the dictionary websites. For example, combinations of grammatical words often appear in the collocation dictionary sections of the websites. The theoretical debate about the difference between collocation and grammar pattern is irrelevant for most dictionary users. However, information about the lexical items which co-occur with a particular wordform, and their syntactic configuration is useful for writers.

By aggregating the number of entries with writing support features such as examples, grammar patterns, and collocation information and dividing this by the total number of items from the sample covered, a *writing support score* can be calculated. This score gives an approximation of how well a resource supports users with the sample items when writing.

$$R = \frac{E_x + G_x + C_x}{N}$$

The overall writing support score is the ratio (R) of the sum of items with examples (E), grammar patterns (G), and collocations (C) for the sample items covered by the dictionary (N). To reflect the diversity in syntactic behaviour of word classes and the varying degrees of difficulty they could present writers, four writing support scores are reported:

— Overall score
—- Score with adverbs excluded
—- Score with adjectives excluded
—- Score with adverbs and adjectives excluded

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1947 (Article)

To calculate the exclusive scores, the sum for items of the included word classes (T) are first weighted (W) representing their proportion of the total sample items covered by the dictionary (N):

$$W_c = \frac{T_c}{N}$$

The exclusive score, a ratio, is then calculated using this weighting.

$$r = \frac{(T_x * W_c) + G_x + C_x}{N}$$

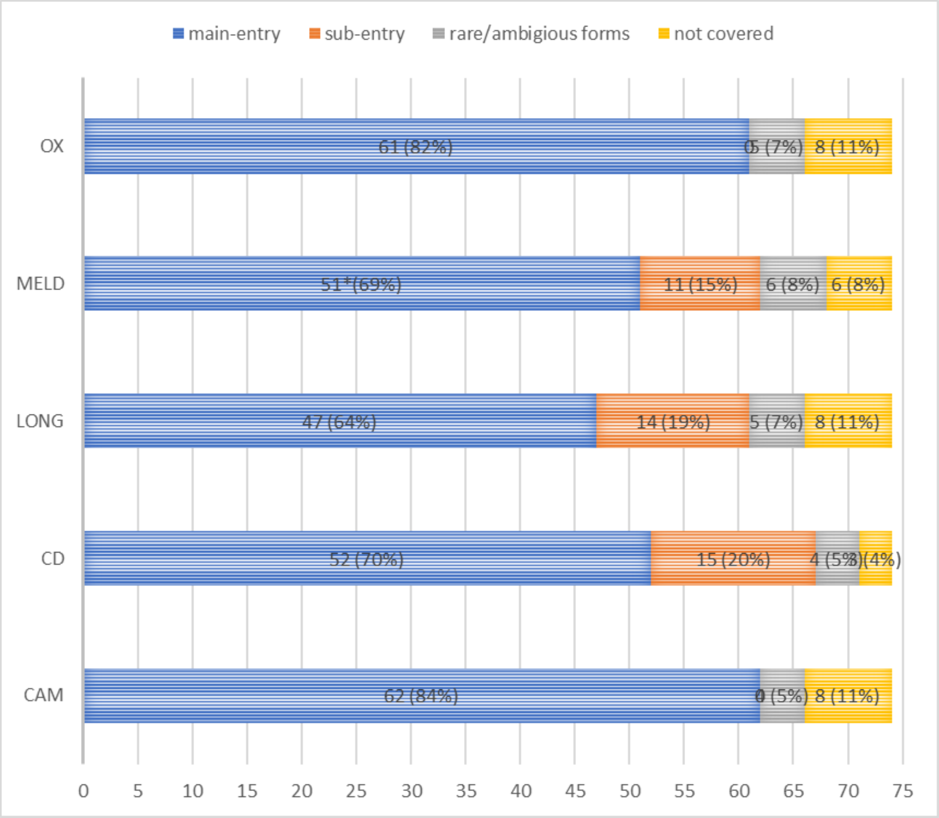
The coverage statistics and writing support scores indicate how well users of the dictionary websites are supported when seeking to use the problematic derivative forms in writing. For a more detailed impression, it is necessary to examine which items have writing support features.

Although examples can provide information about grammar patterns, here analysis focuses on semantics. Namely, whether derived wordforms missing examples are sufficiently semantically regular for a user to infer their meaning and use. This study does not differentiate between exemplification styles employed in the dictionaries. However, it is noteworthy that CAM, CD, and LONG occasionally present examples automatically extracted from corpus lines. When relevant to the target word, these are counted.

Comparing items with and without grammar pattern and collocation information by word class across the dictionary websites provides a clearer impression of how well users are supported when writing the problematic forms. Although users can intuit collocation and grammar patterns from examples, only those instances where the dictionary compiler intentionally highlights these aspects are considered. Common strategies include presenting salient collocations or grammar patterns in bold in examples (all dictionary websites examined) and/or separating common collocates with slashes (e.g., LONG (Figure 3) and OX) and displaying information from the publisher's collocation dictionary for certain searches. Additionally, LONG occasionally provides links to fuller entries for salient collocations and grammar patterns; CD, MELD, and OX display common idioms for some of the sample, while CD includes COBUILD grammar patterns.

Results and discussion

Coverage. The impression of inconsistent treatment of derivative forms reported in previous research is not immediately supported. Most items in the sample are covered by the five websites. The mean number of items treated per website ($N = 74$) is 66.4 with a standard deviation of 3.64. The overall coverage of the sample items did not differ significantly by dictionary website, $X^2 = 3.194$; df = 4; $p < .05$.



*Includes entry for Macmillan OPEN DICTIONARY for *persistency*

Figure 5: Coverage and entry status of problematic wordform on the "Big five" dictionary websites

A high degree of coverage was expected, the sampling criteria ensured target items were used reasonably frequently and widely. Indeed, inclusion in a dictionary was one of the three criteria Schmitt and Zimmerman (2002) used to select the permissible responses to their gapped sentence exercise.

Greater difference is apparent in *how* words are treated. OX and CAM cover all sample items as main entries, while CD, LONG, and MELD use sub entries for around one-sixth of the items. This suggests that although CAM and MELD cover a greater number of items overall, OX and CAM provide better writing support than the other resources. Further analysis of the entry contents is needed to substantiate this.

Table 4: Wordforms which are missing from at least one of the dictionary websites

	CAM	CD	LONG	MELD	OX
accessibly (Level 3)	x	sub	sub	x	x
authoritive (Level 6)	x	x	x	x	x
authoritively (Level 3)	x	x	x	x	x
coherency (Level 6)	x	redirect	redirect	x	x
ethnicity (Level 4)	main	main	x	main	main
liberalness (Level 3)	x	sub	x	x	x
minimization (Level 4)	main	sub	x	sub	main
persistency (Level 4)	x	redirect	x	open	
philosophic (Level 6)	placeholder	main	redirect	redirect	redirect
preciseness (Level 3)	main	sub	x	sub	x
traditionize (Level 4)	x	x	x	x	x

Table 4 shows the eleven wordforms which are missing from at least one website. Only three items are absent from all websites: *authoritive* and *authoritively,* infrequent spellings of *authoritative* and *authoritatively*, and *traditionize* a rarer verb meaning 'to make into a tradition'.

The treatment of *coherency,* a more infrequent form of *coherence,* and *philosophic,* a more infrequent form of *philosophical,* is inconsistent. The former is absent from CAM, MELD, and OX, the latter not found in CAM. Except for *philosophic* in CAM and CD, searching for these wordforms redirects the user to the page for the more frequent form. Once there, the infrequent form is listed after "also" (LONG and OX) or "or" (MELD). The first entry when searching for *philosophic* on CD is a COBUILD entry stating: "*Philosophic* means the same as *philosophical*" with a hyperlink to *philosophical*. Since both wordforms are wholly interchangeable, this redirection strategy seems sound. For resources where the infrequent forms are not listed, the alphabetic proximity of these items to their counterparts means that users may select the relevant form from the alphabetical listing presented when a search produces no exact results. Searching for *philosophic* in CAM produced a placeholder consisting solely of corpus lines for *philosophic.*

The treatment of forms with the Level 3 affix *-ness, liberalness* (only present in CD) and *preciseness* (absent from LONG and OX), may be inconsistent. For example, *preciseness* is in CAM but not *liberalness*. It may be that *liberalness* was considered too infrequent for inclusion[6] or its inclusion may be an oversight given the productivity (almost any adjective + *-ness* produces an acceptable noun) and semantic regularity (meaning "'property of being X', where X is the base adjective" (Carstairs-McCarthy 2018: 78)) of this suffix. However, as these are the only two *-ness* forms in the sample, care must be taken not to overgeneralise.

The omission from CAM, MELD, and OX of *accessibly*, an adverb formed with the Level 3 affix *-ly* could suggest inconsistent coverage. However, the presence of the thirteen other *-ly* adverbs from the sample suggests another factor, possibly frequency, plays a role.

The wordforms *ethnicity* and *minimization* are notably absent from LONG. There are four other occurrences of *-ity*, and two other occurrences of *-ation* sample wordforms covered by the website. Since words formed with *-ity* often have a specialised meaning which "may be hard to deduce" (Bauer and Nation 1993: 275), the omission of *ethnicity* is unfortunate. The omission of *minimization* here is surprising given its frequent semi-technical uses. While the omission of these words formed with often challenging Level 4 affixes could be a simple mistake, it may still inconvenience users.

Beyond coverage, there is less consistency in the way sample items are treated across the websites. One source of confusion is the ambiguous status of *-ed* and *-ing* forms which can be analysed as either adjectives or participle forms and in the case of *-ing* also as nouns. Schmitt and Zimmerman (2002) label the *-ed* forms (*accessed, assumed, authorized, released,* and *selected*) and the *-ing* form (*surviving*) as adjectives.

Table 5: Treatment of ambiguous word class forms

	CAM	CD	LONG	MELD	OX
accessed	verb	verb	verb	verb	verb
assumed	verb	adjective	verb	verb	adjective
authorized	adjective	adjective	adjective	verb	adjective
released	verb	verb	verb	verb	verb
selected	verb	verb	adjective	verb	verb
surviving	adjective	adjective	adjective	adjective	verb

Table 5 shows searches for these *-ed* and *-ing* forms give inconsistent results. All sites redirect searches for the items *accessed* and *released* to *access and release* (v). No adjectival senses of these items are given. The adjective sense of *selected* is a sub-entry of the verbal sense from the LBD. The adjective a*ssumed* is listed as a main entry in CD and OX. The adjective *authorized* is present as a main entry in all the dictionaries except MELD. The adjective *surviving* is present as an entry or sub-entry in all dictionaries except OX. However, there are examples and collocations for the verbal entry which could be analysed as adjectival.

Some of these deficiencies are mitigated, intentionally or otherwise, by fea-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1947 (Article)

tures of online dictionaries. Problems with corpus methods in lexicography often stem from inaccuracies in part-of-speech tagging (Frankenberg-Garcia, Rees and Lew 2021). Many methods tend to treat *-ed* forms as verbs rather than adjectives. This may explain the tendency to treat these forms as participles in the dictionaries. However, it also means that some of the corpus-derived examples in verbal entries could be analysed as adjectives. For instance, the example provided for the fifth sense of *release* (v) in OX: "The newly *released files* reveal […]". This is more apparent still in automatically retrieved examples from corpora. For example, in the entry for *release* in CAM: "To what extent the rural sector absorbs the *released labour* is not clear" and "There are only a few landraces and very old *released varieties* available."

Helpful features include the alphabetical index adjacent to entries on all websites except LONG. For example, on CAM's page for **assume**, the user is presented with adjectival uses: *assumed debt*, *assumed liabilities*, *assumed name* in the 'Browse' box at the bottom of the entry. For years, liberation from the constraints of the alphabetical index has been regarded positively (c.f., De Schryver 2003). However, this feature can mitigate a methodological deficiency in electronic lexicography. Predictive text searches also help users find adjectival senses. For example, in MELD typing *assumed* predicts *assumed name* which is listed as a discrete entry. MELD also contains a crowd-sourced example containing an adjectival use of *authorized*, *authorized push payment*. An example of a crowd-sourced element potentially resolving a deficiency, albeit a relatively minor one, in a professionally produced dictionary.

The prevalence of homographs in English is problematic for electronic lexicography. Table 6 indicates the word class initially displayed when searching for a homographic item. Dictionary search engines cannot determine the user's intended word class. The basic form's ordering might reflect the compilers' view of the primary form or merely the frequency of word classes in the corpora used.

Table 6: Word class first presented for homographic pair

	CAM	CD	LONG	MELD	OX
access (n) vs. *access* (v)	noun	noun	noun	verb	noun
minimum (n) vs. *minimum* (adj.)	noun	noun	adjective	adjective	adverb
release (n) vs. *release* (v.)	verb	verb	verb	verb	verb
select (v) vs. *select* (adj.)	verb	verb	verb	verb	verb

This coverage analysis provides insights into how members of derivationally related WFs are treated in online monolingual English dictionaries. Overall coverage statistics suggest reasonably consistent treatment of the WF members sampled. Inconsistencies include: the omission of forms with the morpheme *-ness* (*liberalness* and *preciseness*) which could be justified by its formal and semantic regularity, inconsistent treatment of rare wordforms which have more frequent equivalents (*coherency* and *philosophic*), and the ambiguous word class of *-ed* and *-ing* wordforms. These minor inconsistencies may not have an impact on the user. Furthermore, electronic lexicography methods both contribute to and mitigate such inconsistencies.

Writing support. The coverage analysis above suggests that members of derivationally related WFs are well covered on the websites examined (RQ1). However, to establish the extent to which they are treated in a way which facilitates productive use in writing (RQ2) a finer-grained analysis is necessary. A key assumption here is that examples, grammar patterns, and collocation information help writers. Another assumption is that the six rare forms with more frequent counterparts can be disregarded since it is likely that users will look up the more frequent counterpart.

Table 7: Writing support scores

	CAM	CD	LONG	MELD	OX
Overall score	1.9	2.1	1.8	1.5	2.1
Score with adverbs excluded	2.2	2.1	1.9	1.8	2.3
Score with adjectives excluded	2.0	1.9	1.7	1.6	2.1
Score with adverbs and adjectives excluded	1.8	1.7	1.5	1.4	1.9

The writing support scores in Table 7 suggest that OX provides the most comprehensive writing support for the problematic wordforms, closely followed by CAM and CD. MELD's score is notably lower than the others. This relation holds for the exclusive scores. However, caution is needed when interpreting differences in such a small sample. These scores indicate inconsistency in *how* the sample is treated across the websites examined. A closer examination of the individual components of writing support (examples, grammar patterns, and collocation information) confirms this impression and elucidates differences in sample treatment within dictionary websites.

The proportion of items with examples (Figure 6) differs significantly by website, $X^2 = 30.068$; df = 4; $p < 0.001$. Both OX and CAM provide examples for 97% of the items they cover. Items missing examples are the ambiguous word class forms *accessed* (CAM and OX), *assumed* (CAM), and *selected* (OX). Since the dictionaries treat them as verbs and provide examples for the verbal senses, they effectively offer examples for all items they list. CD provides examples for 88% of items covered. Again, two ambiguous class items (*accessed* and *released*) lack examples. LONG provides examples for 86% of items it covers including *accessed, assumed,* and *released*. The outlier here is MELD where 70% of items covered have examples.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1947 (Article)

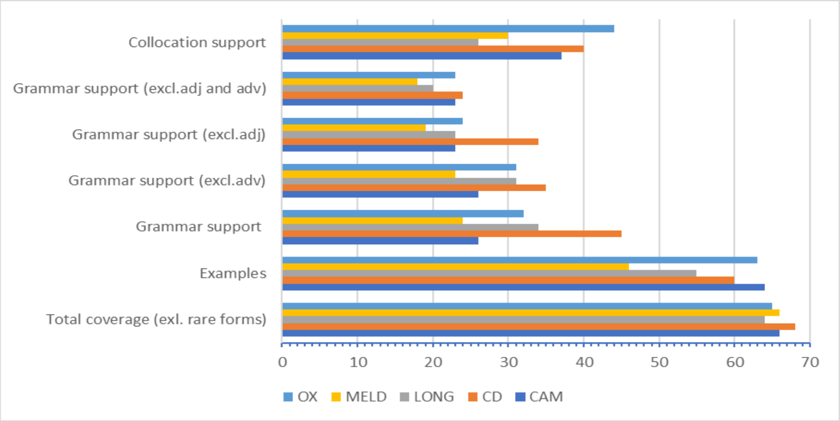


Figure 6: Sample items with writing support features per resource

There is clear inconsistency in the provision of grammar patterns on the websites examined (Figure 6). The proportion of items with this information differs significantly by website, $X^2 = 14.2796$; df = 4; $p < 0.006$. Overall, CD leads providing information for 60% of items covered. LONG provides grammar information for 53% of items covered, followed by OX (49%). CAM provides grammatical information for 39% of covered items, MELD for 36%. The syntactic behaviour of different parts-of-speech poses different degrees of challenge for writers. However, this trend persists when adverbs are excluded. For example, with a coverage statistic of 72% CD is notably higher than OX (55%), LONG (53%), and CAM (51%), and considerably more so than MELD (42%). When adjectives and adverbs are excluded, OX has the highest statistic (72%) followed by CD (71%) and CAM (70%); LONG covers 67% of noun and verb items, with MELD lower at 55%.

The proportion of items with collocation information (Figure 6) differs significantly by website, $X^2 = 12.192$; df = 4; $p < 0.05$. OX leads by providing collocation information for 68% of items covered, followed by CD (59%) and CAM (56%), then MELD (45%), and finally LONG (41%).

Since many users can induce information about grammatical patterns and collocational behaviour from dictionary examples and corpus lines, the relative absence of grammar patterns on CAM and OX is perhaps mitigated by their comprehensive example provision. This is reflected in the overall writing support score.

From the broad view adopted so far, considerable variation in the provision of writing support features between dictionaries is apparent. The following three sub-sections provide a finer-grained analysis of this variation.

Examples. As Table 8 indicates, after ambiguous word class forms, *-ly* adverbs are the wordforms most frequently missing examples. In general, they are semantically regular "Xly means 'in an X fashion' for any adjective X" (Carstairs-McCarthy 2018: 20). This general rule applies to *coherently*, *ethnically*, *minimally*, and *persistently* (all lacking examples in MELD). However, *accessibly*, *authoritatively*, *ideologically*, and *philosophically* are edge cases. For instance, without an example learners lacking deep relational knowledge could conceivably make the erroneous connection *philosophy* → *philosophical* ('related to philosophy') → *philosophically* (in a 'manner related to philosophy') rather than the prototypical meaning "in a way that calmly accepts a difficult situation" (CAM).

Table 8: Items missing examples

	CAM	CD	LONG	MELD	OX
accessed	X	X	X	X	X
accessibility			X		
accessibly	NC	X	X	NC	NC
assumed	X		X	X	
authoritatively			X	X	
authorized				X	
cohere				X	
coherence				X	
coherently				X	
ethnically				X	
ethnicity			NC	X	
ideologically		X	X		
liberality				X	
liberalization				X	
liberalize				X	
liberalness	NC	X	NC	NC	NC
minimally				X	
minimization		X	NC	X	
persistently				X	
philosophically		X	X	X	
philosophize				X	
preciseness		X	NC	X	NC
released		X	X	X	
selected				X	X
selectively			X		

Note. NC = not covered by dictionary website; X = missing

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1947 (Article)

The high degree of productivity and semantic regularity of the affix *-ness* which generally means "'property of being X', where X is the base adjective." (Carstairs-McCarthy 2018: 78) could explain the omission of *liberalness* from all resources except CD and *preciseness* from LONG and OX, and the omission of an example for *liberalness* (CD) and *preciseness* (CD and MELD). However, the presence of examples for these items in the other resources suggest their creators do not share this assumption of relational knowledge.

The lack of examples for *liberalization* (MELD) and *minimization* (CD and MELD) can be explained by the generalizability of *-ation*. However, as with the absence of *minimization* from LONG, both wordforms have a specialised meaning frequent in academic context (e.g., "He is a longtime proponent of his country's economic liberalisation." (CAM); "cost minimization" (OX)). The absence of *liberalize* from MELD is notable for the same reason (e.g., "They will work with a view to further liberalize the investment regime" (CAM)). Like the absence of an example for *philosophically* discussed in the coverage analysis above, the absence of an example for *philosophize* from MELD is problematic as it does not typically mean 'to create philosophy' rather "to talk for a long time about subjects such as the meaning of life" (CAM). An example could also demonstrate that, in contrast to many words derived with the affix *-ize,* it is intransitive. The following examples from CAM for the entries for the *-ize* forms sampled illustrate complementation patterns well:

I authorized my bank to pay her £3,000.

They have plans to liberalize the prison system.

We must minimize the risk of infection.

Students, she complained, had nothing better to do than spend whole days philosophizing about the nature of truth.

The provision of examples for wordforms derived with *-ity* is also problematic. The missing example for *accessibility* in LONG is surprising. Firstly, because examples exist in the other dictionaries and, secondly, because it has a specialised yet frequent sense: "how easy something is to reach, enter, use, etc. for somebody with a disability" (OX). Additionally, examples for *ethnicity* and *liberality* are missing from MELD*.* As Schmitt and Zimmerman (2002) show, the extent to which productive knowledge of these words is easily predictable from productive knowledge about their base is questionable. When the base has two or more senses this assumption of relational knowledge entails a further assumption: that the user knows which sense is relevant to the derivative. For example, the definitions below come from CAM: (1) and (2) define *ethnic*, (3) defines *ethnicity*. The relation between (1) and (3) is immediately apparent. The relation between (2) and (3) requires some mental gymnastics.

1. *relating or belonging to a group of people who can be seen as distinct (= different) because they have a shared culture, tradition, language, history, etc.:*

2. *seen as different or interesting because of coming from a culture or tradition that is not Western:*
3. *a large group of people with a shared culture, language, history, set of traditions, etc., or the fact of belonging to one of these groups:*

Examples for the WF members *coherence* and *cohere* are notable omissions from MELD. This may stem from an assumption that learners have the relational knowledge to make the connection to the adjective *coherent*. This is particularly questionable in the case of *coherence* as although the Level 5 affix *-ence* is reasonably regular, it is not frequent (Bauer and Nation 1993: 260).

Regarding the provision of examples, there is clear inconsistent treatment across dictionaries, and in the case of CD, LONG and MELD, within dictionaries. Barring the ambiguous word class items, in CD and LONG the sub-entry status of items may be an explanatory factor for, or a consequence of, the missing examples. However, in MELD both main- and sub-entries lack examples.

Grammar patterns. A comparison of items with (Figure 7) and without (Figure 8) grammar pattern information suggests inconsistent writing support between and within websites.

As discussed, the need for grammar pattern information varies by word class. All sampled adverbs given grammatical support in CD have main entry status. Their grammar patterns come from COBUILD. Many first appeared in the 'extra-column' of the paper dictionary (Hands 2018) and were migrated online. Wordforms lacking grammar pattern support occur as "derived words" and sub-entries in other Collins dictionaries such as *Collins English Dictionary* and *Webster's New World College Dictionary*. Similarly, entry status explains the presence of grammatical information for adverbs in LONG. Those with support are the "Sentence adverbs", *inevitably* and *traditionally*, and *precisely.* The latter is followed by the interrogative pronouns *how/when/where*. All adverbs lacking grammar support in LONG, except *liberally,* are sub-entries. In MELD, *precisely* is also listed followed by *how/when/what* and in OX it followed by *because.* All other sampled adverbs in the latter two dictionaries lack grammatical pattern information.

Adjectives selecting prepositions (*accessible* ***to***, *liberal* ***with***, *minimum* ***of*** etc.) are treated fairly consistently. Inconsistencies occur in CD, LONG, and OX, which mark typical word order for some adjectives (e.g., "precise [adj NOUN]") but not others with the same order (e.g., *coherent*). Dictionaries that do not indicate this order (e.g., CAM and MELD) offer less detailed yet more consistent treatment.

All sampled verbs in CD have grammar pattern information. Patterns for *cohere* are absent from CAM, LONG, and MELD. Of the *-ize* affixed verbs, only *liberalize* has pattern information in CD, while patterns for *philosophize* are absent in LONG, MELD, and OX. As discussed, grammar pattern information may be useful for learners wishing to use *philosophize* as it is a rare example of an intransitive verb derived with *-ize* which frequently occurs with the prepositions *of* or *about*, as documented in CAM and CD. Similarly, *cohere with* is a typical pattern given in OX and CD.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1947 (Article)

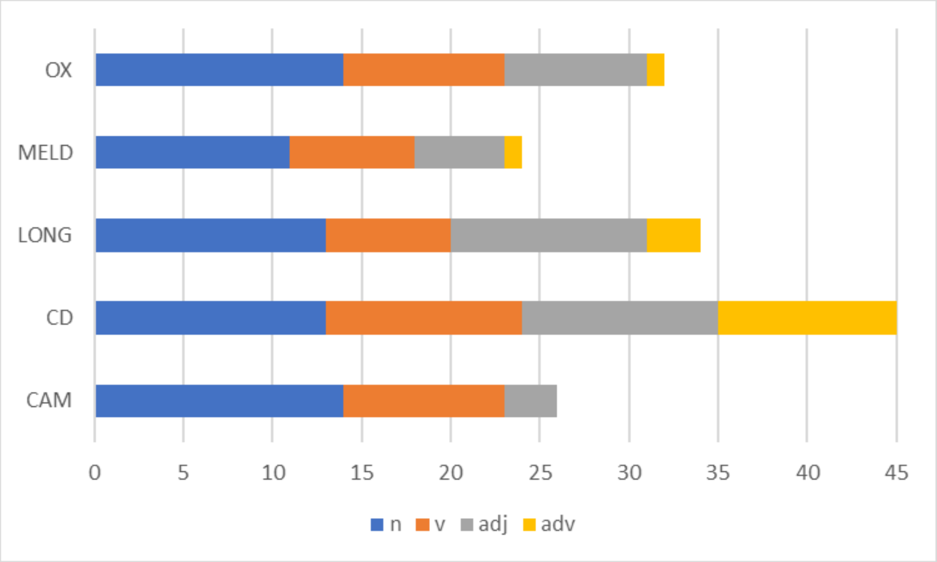


Figure 7: Sample items with grammar pattern information per resource

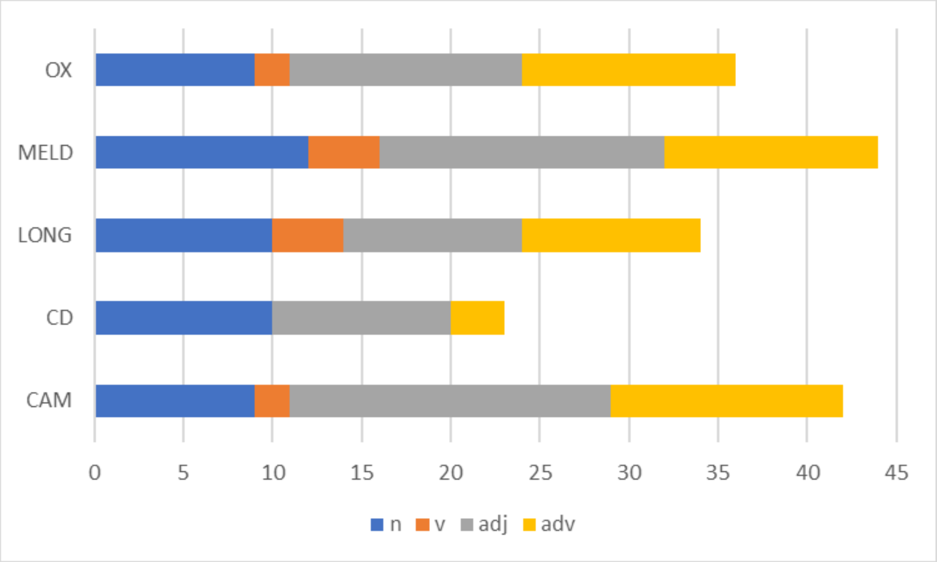


Figure 8: Sample items without grammar pattern information per resource

Nouns are derived using a greater variety of affixes than other word classes. Table 9 shows the sample nouns included on each website and whether they have grammar pattern information. The overall impression is one of inconsistent treatment within and between dictionaries.

Table 9: Nouns with grammar pattern information

	Without grammar patterns	Words with grammar patterns
CAM	*ethnicity* *ideology* *inevitability* *liberality* *liberalization* *preciseness* *precision* *survival** (The phrase *survival of the fittest* is listed)	*access* *accessibility* *assumption* *authority* *authorization* *coherence* *liberty* *minimization* *minimize* *minimum* *persistence* *philosophy* *release* *selection* *tradition*
CD	*authority* *authorization* *coherence* *ethnicity* *ideology* *liberality* *liberalness (sub)* *minimization (sub)* *preciseness (sub)* *survival** (The phrase *survival of the fittest* is listed)	*access* *accessibility* *assumption* *inevitability* *liberalization* *liberty* *minimize* *minimum* *persistence* *philosophy* *precision* *release* *selection* *tradition*
LONG	*accessibility* (sub) *authorization* *coherence* *ideology* *liberality* *liberalization* (sub)	*access* *assumption* *authority* *inevitability* *liberty* *minimum*

		persistence *philosophy* *precision* *release* *selection* *survival* *tradition*
MELD	*accessibility* (sub) *coherence* *ethnicity* *ideology* *liberality* *liberalization* (sub) *liberalness* *minimization* (sub) *minimize* *persistence* *philosophy* *preciseness* (sub) *precision*	*access* *assumption* *authority* *authorization* *inevitability* *liberty* *minimum* *release* *selection* *survival* *tradition*
OX	*ethnicity* *ideology* *inevitability* *liberality* *liberalization* *minimization* *persistence*	*access (noun)* *accessibility* *assumption* *authority* *authorization* *coherence* *liberty* *minimize* *minimum* *philosophy* *precision* *release* *selection* *survival* *tradition*

Wordforms without grammar pattern information are predominantly derived by affixation using *-ity* and *-ation*. Those that do have grammar patterns can be analysed as the base wordforms or are often the most frequent member of their family according to Schmitt and Zimmerman's (2002) counts. The usefulness of grammar pattern information for these items to writers can only be ascertained by direct empirical research. However, it is notable that producing these wordforms posed problems for Schmitt and Zimmerman's (2002) participants.

Some items missing grammatical patterns exhibit similar grammatical behaviour to those which have them. For example, *assumption that* appears in all resources while *inevitability that* is absent form CAM and MELD. This suggests a need for grammatical pattern information for many items missing it. Like the provision of examples, many of the wordforms without grammatical pattern information were treated as subentries, irrespective of their word class.

Collocation information. Unlike closed classes or phrasal categories that constitute grammar patterns, the range of potential collocates is limitless. Variation in typical collocates presented for a given base between resources is expected due to variation in corpus composition. Consequently, this analysis of collocation information must adopt a broad focus.

The provision of collocation information does not follow the general trend for writing support in the dictionaries examined. Notably, LONG rather than MELD provides collocation information for fewest items. However, differences exist across word classes.

Table 10: Noun items missing collocation information

	CAM	CD	LONG	MELD	OX
accessibility		X	X	X	X
authorization			X		
coherence			X		
ethnicity	X	X	NC	X	X
ideology			X		
inevitability	X			X	X
liberality		X	X	X	X
liberalization		X	X	X	X
liberalness	X	X	NC	X	NC
minimization		X	NC	X	X
minimum		X			
persistence	X			X	
philosophy				X	
preciseness	X	X	NC	X	NC
precision				X	
selection		X			

Note. NC = not covered by dictionary website; X = missing

Collocation information is absent for three out of twenty-three noun items (Table 10) in all resources: *ethnicity*, *liberalness*, and *preciseness*. Five items (*accessibility*, *liberality*, *liberalization*, and *minimization* in CAM; and *inevitability* in COD)

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1947 (Article)

only have it in one resource. In contrast, seven items are absent from one resource (*minimum* and *selection* from COD; *authorization*, *coherence*, *ideology* from LONG; and *philosophy* and *precision* from MELD).

Table 11: Verb items missing collocation information

	CAM	CD	LONG	MELD	OX
access	X			X	
assume	X				
authorize			X	X	
cohere		X	X	X	X
liberalize		X	X	X	
minimize	X	X			X
persist			X		
philosophize		X	X	X	X
release	X				
select	X		X		
survive			X		

Note. X = missing

All eleven verb items have collocation information in at least one resource (Table 11). Although, for *cohere* and *philosophize,* this information is only provided by CAM. This is problematic because it assumes relational knowledge with other family members. Three resources lack information for *liberalize* (CD, LONG, and MELD) and *minimize* (CAM, CD, and OX). As with examples, some academic writers might benefit from collocation information about these semi-technical terms.

Table 12: Adjective items missing collocation information

	CAM	CD	LONG	MELD	OX
accessed	X	X	X	X	X
assumed	X		X	X	
authoritative			X	X	X
authorized	X		X	X	X
coherent	X		X		
ideological	X		X	X	

liberal				X	
minimal			X		
minimum		X			
philosophical			X	X	
precise			X		
released	X	X	X	X	
select				X	
selected	X	X	X	X	
selective			X		
surviving	X	X		X	

Note. X = missing

Ostensibly, provision of collocation information for adjectives is less comprehensive than for nouns and verbs (Table 12). However, seven of the items missing collocation information are ambiguous word class items treated as verbs. Moreover, three resources lack information for *ideological* (CAM, LONG, and MELD) three for *authoritative* (LONG, MELD, and OX), two for *coherent* (CAM and LONG) and *philosophical* (LONG and MELD).

Table 13: Adverb items missing collocation information

	CAM	CD	LONG	MELD	OX
accessibly	NC	X	X	NC	X
authoritatively	X	X	X	X	
coherently	X	X	X	X	X
ethnically	X	X	X		X
ideologically	X	X	X	X	X
inevitably	X	X	X	X	X
liberally	X		X	X	X
minimally		X	X	X	X
persistently	X		X	X	X
philosophically	X	X	X	X	X
precisely	X				
selectively	X	X	X	X	
traditionally	X	X	X	X	X

Note. NC = not covered by dictionary website; X = missing

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1947 (Article)

Provision of collocation information for adverbs is the least comprehensive of all word classes (Table 13). Information is provided for *precisely* in all resources except CAM. CD also provides information for *liberally* and *persistently,* OX for *authoritatively* and *selectively,* and CAM for *minimally*. Two factors may explain this sparse coverage: Firstly, the suffix *-ly* is extremely semantically regular "Xly means 'in an X fashion', for any adjective X." (Carstairs-McCarthy 2018: 20), so presumably lexicographers assume users can use the *-ly* adverbs in production by connecting them to their knowledge of the adjective base. Secondly, users are unlikely to start a collocation search using an adverb: "It would not make sense for a writer to initiate a collocation query from an adverb (e.g. 'what words can I use with *primarily?*')" (Frankenberg-Garcia et al. 2019: 28).

This analysis of grammar support features for derivative forms suggests examples, grammar patterns, and collocations work independently when supporting writers. This is unlikely; writers may take information simultaneously from all three sources. If one feature (e.g., grammar pattern information) is unavailable they may rely more heavily on another (e.g., examples). Future analysis of writing support would benefit from a model reflecting this relationship.

Conclusions

This study aimed to investigate the treatment of academic WFs on five English dictionary websites frequently used by learners. It was motivated by a belief that the members of these WFs should be treated in a way that facilitates learners' written production. Two factors prompted this belief: Firstly, research demonstrating that when given a basic prompt wordform, academic writers struggle producing derivative forms from the same WF. Secondly, the removal of space restraints in electronic resources, which hypothetically allows more detailed coverage of derivatives than paper-based dictionaries.

Overall, the five websites examined cover most items in the sample of challenging wordforms. This good coverage contrasts with findings on paper-based dictionaries. However, as in previous research, there is considerable variation in the treatment of derivative wordforms within and between resources.

The quantity of writing support features varies greatly across websites. Although MELD covers a high proportion of sample items, it provides fewer examples, grammar patterns, and collocation information than the other resources. Within resources, the reasons for inclusion or exclusion of items and their related writing support features are not always clear. For certain affixes, this may be due to assumptions about generalisability of their semantic or syntactic behaviour. These assumptions may be misguided since empirical research suggests writers do not always connect bases and derivatives formed by suffixation even with highly generalisable and productive affixes. Occasionally, (e.g., *ethnicity, liberalization*), analysis of the excluded wordforms suggests their semantic relationship to the base is idiosyncratic. Alternatively, their relative frequency in corpora used in compilation may explain exclusion. Further investigation here would be beneficial.

Further research could also mitigate limitations restricting the generalisability of these conclusions. Important limitations relate to the 74 problematic wordforms investigated. Not only is this sample small, but its items are also morphologically limited containing a relatively narrow range of suffixes. Future research should investigate forms created via prefixation (e.g., with *co-*, *in-*, *re-* etc.) if producing these is found to be a problem for writers.

Practical considerations for dictionary makers. Assumptions about users' relational knowledge of WF members should be reevaluated. Instead of assuming that writers can connect the base, the affix and derivative meaning, dictionary makers should aim for more complete treatment of derivatives. Electronic resources, unrestrained by the physical restrictions of paper-based dictionaries, could offer users fuller entries for derivative forms. However, compiling dictionary entries costs money. Deprived of income from sales of paper dictionaries, it is unlikely that publishers will invest in this. Nonetheless, as seen with corpus lines and collocation lists, methods from electronic lexicography can, sometimes inadvertently, offer a solution.

Endnotes

1. For a more nuanced view, see Lew (2011) who makes a distinction between the potentially unlimited *storage space* for lexicographic data and more limited *presentation space* on the user's screen.
2. *Macmillan English Dictionary online* was shut down on June 30th, 2023.
3. The edition of the LBD from which the entry is taken is not specified.
4. The editions of COBUILD and CED from which the entries are taken are not specified.
5. The dictionaries mentioned are *Cambridge International Dictionary of English* (Procter 1995), *COBUILD English Learner's Dictionary* (Sinclair 1989), *Longman Dictionary of English Language and Culture* (Summers 1992), and *Oxford Advanced Learner's Dictionary of Current English* (Crowther 1995).
6. This would be surprising; "we checked the frequency of these derivatives in the BNC and considered eliminating those that had very low frequency counts or did not exist in the corpus." (Schmitt and Zimmerman 2002: 156)

References

Agnes, M. (Ed.). 2010. *Webster's New World College Dictionary*. Fourth Edition. Cleveland, Ohio: Wiley.

Bauer, L. and P. Nation. 1993. Word Families. *International Journal of Lexicography* 6(4): 253-279.

Berko, J. 1958. The Child's Learning of English Morphology. *Word* 14(2–3): 150-177.

Carstairs-McCarthy, A. 2018. *An Introduction to English Morphology: Words and Their Structure*. Kindle Edition. Edinburgh: Edinburgh University Press.

Coxhead, A. 2000. A New Academic Word list. *TESOL Quarterly* 34(2): 213-238.

Crowther, J. (Ed.). 1995. *Oxford Advanced Learner's Dictionary of Current English*. Fifth Edition. Oxford: Oxford University Press.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1947 (Article)

De Caluwe, J. and J. Taeldeman. 2003. Morphology in Dictionaries. Van Sterkenburg, P. (Ed.). 2003. *A Practical Guide to Lexicography*: 114-126. Amsterdam: John Benjamins.

De Schryver, G.-M. 2003. Lexicographers' Dreams in the Electronic-Dictionary Age. *International Journal of Lexicography* 16(2): 143-199.

Frankenberg-Garcia, A. 2015. Dictionaries and Encoding Examples to Support Language Production. *International Journal of Lexicography* 28(4): 490-512.

Frankenberg-Garcia, A., R. Lew, J.C. Roberts, G.P. Rees, and N. Sharma. 2019. Developing a Writing Assistant to Help EAP Writers with Collocations in Real Time. *ReCALL* 31(1): 23-39.

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.

Hands, P. 2018. COBUILD Design and Layout: Changes over the Last 30 Years. *Collins Dictionary Language Blog*.
https://blog.collinsdictionary.com/language-lovers/cobuild-design-and-layout-changes-over-the-last-30-years/ [30 July 2023]

Hanks, P. 2013. *Lexical Analysis: Norms and Exploitations*. Cambridge, MA: MIT Press.

Hyland, K. and Polly Tse. 2007. Is There an "Academic Vocabulary"? *TESOL Quarterly* 41: 235-253.

Kilgarriff, Adam, Miloš Husák, Katy McAdam, Michael Rundell and Pavel Rychlý. 2008. GDEX: Automatically Finding Good Dictionary Examples in a Corpus. Bernal, Elisenda and Janet 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.

Lew, R. 2011. Space Restrictions in Paper and Electronic Dictionaries and their Implications for the Design of Production Dictionaries. Bański, Piotr and Beata Wójtowicz (Eds.). 2011. *Issues in Modern Lexicography*. München: Lincom Europa.

Procter, P. (Ed.). 1995. *Cambridge International Dictionary of English.* Cambridge: Cambridge University Press.

Ptasznik, B. 2023. More Examples May Benefit Dictionary Users. *International Journal of Lexicography* 36(1): 29-55.

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.

Schmitt, N. 1998. Tracking the Incremental Acquisition of Second Language Vocabulary: A Longitudinal Study. *Language Learning* 48(2): 281-317.

Schmitt, N. 1999. The Relationship between TOEFL Vocabulary Items and Meaning, Association, Collocation and Word-class Knowledge. *Language Testing* 16(2): 189-216.

Schmitt, N. and P. Meara. 1997. Researching Vocabulary through a Word Knowledge Framework: Word Associations and Verbal Suffixes. *Studies in Second Language Acquisition* 19(1): 17-36.

Schmitt, N. and C. Zimmerman. 2002. Derivative Word Forms: What Do Learners Know? *TESOL Quarterly* 36(2): 145-171.

Sinclair, J. (Ed.). 1989. *COBUILD English Learner's Dictionary*. London: Collins.

Stein, G. 1985. Word-formation in Modern English Dictionaries. Ilson, R. (Ed.). 1985. *Dictionaries, Lexicography and Language Learning*: 35-44. Oxford: Pergamon.

Summers, D. (Ed.). 1992. *Longman Dictionary of English Language and Culture.* Harlow: Longman.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1947 (Article)

Ten Hacken, P., A. Abel and J. Knapp. 2006. Word Formation in an Electronic Learners' Dictionary: ELDIT. *International Journal of Lexicography* 19(3): 243-256.

Tyler, A. and W. Nagy. 1989. The Acquisition of English Derivational Morphology. *Journal of Memory and Language* 28(6): 649-667.

Van Sterkenburg, P. 1992. *Het Woordenboek der Nederlandsche taal : Portret van een taalmonument.* The Hague: Sdu.

West, M. 1953. *A General Service List of English Words*. London: Longman, Green & Co.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1948 (Article)

The Inclusion of Neologisms in the Revision of the *Grand Dictionnaire Chinois–Français Contemporain*

Fang Huang, *Centre for Linguistics and Applied linguistics, Center for Lexicographical Studies, Guangdong University of Foreign Studies, Guangzhou, China*
(huangfang@gdufs.edu.cn) (https://orcid.org/0009-0009-6947-8364)
and
Jianhua Huang, *Centre for Linguistics and Applied linguistics, Guangdong University of Foreign Studies, Guangzhou, China*
(jhhuang@gdufs.edu.cn) (https://orcid.org/0009-0008-2597-1037)

Abstract: Language change presents continuous challenges for lexicographers, especially with the rapid emergence of neologisms. The *Grand Dictionnaire Chinois–Français Contemporain* (GDCFC) was published in 2014 and its revision began immediately in order to keep the dictionary up to date. A central focus of the revision has been the inclusion of new words and new senses. This article describes the experiences of the dictionary revision team on the inclusion of neologisms using a wide spectrum of sources, ranging from the latest monolingual and bilingual dictionaries, monolingual and bilingual corpora, to online resources. It addresses four major challenges encountered by the team in the ongoing revision process, that is, the inclusion of neologisms absent from Chinese authoritative dictionaries into the GDCFC, the selection of variant neologisms and their French equivalents, the dilemma of whether to include neologisms as entries or merely as examples, and the challenge of ensuring complete relevance between neologisms and the existing entries. In light of these challenges, the article puts forward four criteria, namely, descriptivism complemented with prescriptivism, frequency, supplementation, and relevance. With the recommendations offered herein we hope to provide valuable insights into future lexicographic work on neologisms in bilingual dictionaries.

Keywords: *GRAND DICTIONNAIRE CHINOIS–FRANÇAIS CONTEMPORAIN*, NEOLOGISMS, PARALLEL CORPORA, SOURCES, CHALLENGES, CRITERIA

Opsomming: Die insluiting van neologismes in die hersiening van die *Grand Dictionnaire Chinois–Français Contemporain*. Taalverandering bied, veral met die vinnige ontwikkeling van neologismes, voortdurende uitdagings aan leksikograwe. Die *Grand Dictionnaire Chinois–Français Contemporain* (GDCFC) is in 2014 gepubliseer waarna onmiddellik met die hersiening daarvan begin is om die woordeboek bygewerk te hou. 'n Sentrale fokus van die hersiening was die insluiting van nuwe woorde en nuwe betekenisse. In hierdie artikel word die woordeboekhersieningspan se ervarings rakende die insluiting van neologismes beskryf. Hulle het gebruik

Lexikos 34 (AFRILEX-reeks/series 34: 2024): 469-485

gemaak van 'n wye spektrum bronne wat strek van die jongste eentalige en tweetalige woordeboeke en eentalige en tweetalige korpusse tot aanlyn bronne. Vier hoofuitdagings wat die span tydens die deurlopende hersieningsproses teëgekom het, word bespreek: die insluiting van neologismes in die GDCFC wat ontbreek in Chinese gesaghebbende woordeboeke, die seleksie van variante neologismes en hul Franse ekwivalente, die dilemma rakende die insluiting van neologismes as inskrywings of bloot as voorbeelde, en die uitdaging om algehele relevansie tussen neologismes en bestaande inskrywings te verseker. Met inagneming van hierdie uitdagings word daar in hierdie artikel vier kriteria, naamlik deskriptivisme, aangevul deur preskriptivisme, frekwensie, aanvulling, en relevansie, voorgestel. Met die aanbevelings wat in hierdie artikel gemaak word, word daar gehoop om waardevolle insigte in toekomstige leksikografiese werk rakende neologismes in tweetalige woordeboeke te verskaf.

Sleutelwoorde: *GRAND DICTIONNAIRE CHINOIS–FRANÇAIS CONTEMPORAIN*, NEOLOGISMES, PARALLELLE KORPORA, BRONNE, UITDAGINGS, KRITERIA

1. Introduction

The *Grand Dictionnaire Chinois–Français Contemporain* (GDCFC), chief-edited by Chinese lexicographer, Professor Jianhua Huang, is one of the largest Chinese–French dictionaries. Published in 2014 by Foreign Language Teaching and Research Press, it was "designed and compiled to meet the needs of both Chinese-speaking learners of French and French-speaking learners of Chinese" (Huang and Xu 2019: 325). Cao (2021: 234) commends the dictionary for responding to the needs experienced in the training of Chinese–French professional translators, emphasizing its crucial role in fostering improved communication and cultural exchange across the two languages. In 2018, it was honored with the "Chinese Government Award for Publishing", the highest-level award in the Chinese publishing industry (Huang and Xu 2019: 336).

As Chen (1980: 137) observed, "when a dictionary is published, it is already outdated" because new words and expressions crop up at rapid rate each year. In recent years, China has seen the rapid emergence of new terms that mirror its societal and technological phenomena and advancements. Notable examples include "一带一路" (the Belt and Road Initiative), "微信" (WeChat), "获得感" (sense of gain), "充电宝" (power bank), "自动驾驶" (autonomous driving), "电动汽车" (electric car), "生成式人工智能" (generative artificial intelligence), etc. To keep the dictionary abreast of time and address the evolving needs of users regarding neologisms, the GDCFC revision team has been occupied with updating the dictionary since its publication. Neologisms are defined in this article as recent words or meanings not yet included in the dictionary's first edition.

This article details the team's methodical investigation of neologisms, specifying the challenges encountered and the criteria suggested to tackle these challenges within the GDCFC. It aims to offer valuable perspectives for future research on neologisms in bilingual dictionaries.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1948 (Article)

2. Literature review

Determining how to include a neologism in a dictionary presents significant challenges for lexicographers (Wang 1992; Su and Huang 2003; Yu et al. 2003). Relevant research primarily addresses lexicographers' different attitudes towards the inclusion of neologisms, the challenges associated with neologisms in bilingual dictionaries, and the principles and strategies lexicographers have suggested to tackle the challenges. The following subsections of this literature review highlights how these critical issues have been presented in the existing scholarship.

2.1 Attitudes towards the inclusion of neologisms

Lexicographers' attitudes towards new words vary, representing a spectrum from prescriptive to descriptive stances. This array of perspectives mirrors the ongoing debate over the role of dictionaries.

On the one hand, the prescriptive method is a long-term tradition in lexicography (Balteiro 2011; Gouws and Potgieter 2010). Many dictionaries, particularly early ones, take a prescriptive approach, advocating for guiding language usage among their users. Samuel Johnson's dictionary (1755) confirmed this approach in lexicography (Gouws and Potgieter 2010). Similarly, the *Dictionnaire de l'Académie française* (Walter 2016) and the *Diccionario de la lengua española de la Real Academia Española* (Hanks 2013: 515) also adopt a prescriptive stance. The tradition considers dictionaries as gatekeepers for new words. Based on this perspective, Cheng (2006), taking the *Contemporary Chinese Dictionary* (CCD) as an example, maintains that dictionaries should follow national standards in dealing with the variants, pronunciation of Chinese characters, and the inclusion of scientific terms, in order to further highlight the normative characteristics of dictionaries and enhance their practicality. Some lexicographers also advocate for avoiding the inclusion of new loanwords from foreign languages to maintain the purity of a language (Marello 2020: 176).

On the other hand, following the advent of descriptive linguistics, the field of English lexicography has been predominantly influenced by descriptivist approaches since the 1960s (Wang and Lu 2007: 6). Following this trend, many dictionaries adopt a descriptive synchronic principle, as the *American College Dictionary* (Hanks 2013) and the *English–Chinese Dictionary* (Unabridged) (Wang and Lu 2007) do in their revisions. Wang and Lu (2007: 7) also highlight "the importance of the principle of descriptivism in bilingual-dictionary revision" because it is the principle by which lexicographers "examine and exploit all types of evidence and all sorts of facts and complexities in language use".

Nevertheless, some researchers hold the view that dictionaries are not purely descriptive or prescriptive. Ten Hacken (2020) discusses the contrasting perspectives based on different theoretical frameworks of language. He points out that while lexicographers generally perceive their role as descriptive, docu-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1948 (Article)

menting actual language usage, users often expect dictionaries to function as gatekeepers. Mugglestone (2015: 546) maintains the interaction between the two approaches in dictionary making, claiming that "descriptive processes of collection and evaluation of evidence can be accompanied by prescriptive (and proscriptive) reservation". The combinatory perspective is also adopted in some Chinese–foreign language dictionaries. For example, the *Chinese–English Dictionary* (unabridged) pursues a reserved descriptivism (Lu 2015), primarily describing the actual usage of the language, while not merely recording "anything heard", but instead following a normative principle of selection and elimination for inclusion. As for the GDCFC, it focuses on both approaches when including neologisms. This will be further discussed in detail in 5.1.

2.2 Challenges associated with neologisms in bilingual dictionaries

Scholars in the field of bilingual lexicography, such as Gao (2003), Jin (2007, 2008, 2009), Wang (2010), Zhao (2014, 2015) and Du (2019), have highlighted challenges of including neologisms in bilingual dictionaries.

A prominent challenge is the imbalanced inclusion of neologisms. Zhao (2015: 462) argues that dictionaries typically focus on adding numerous neologisms in the fields of science, technology, and economics, while often overlooking everyday terms, old words with new meanings, and old words with new usage. For instance, the *New Age Chinese–English Dictionary* (2nd edition) fails to include some popular Chinese neologisms like "私房菜" (family cuisine; private kitchen cuisine), "招牌菜" (signature dish or house special), etc. (Zhao 2014: 448).

Another area of focus pertains to the translation of neologisms. Wang (2010) identifies four deficiencies in translating neologisms in Chinese–Japanese dictionaries: improper translation, redundant definitions, absence of translation for implicit components, and neglect of cultural elements. Gao (2003) highlights discrepancies and imprecise definitions for neologisms, while Yang (2014) focuses on excessive translations and inconsistent translation of terminology in political neologisms. Besides, Jin (2008, 2009), Gao (2003), Zhao (2014), and Du (2019) provide specific instances of errors and unidiomatic expressions in the translation of neologisms in bilingual dictionaries. For example, Jin (2009) finds that the term "金球" (golden goal) in the football domain is incorrectly translated as "gold goal". Similarly, the figurative meaning of "充电" is "acquiring new knowledge and skills through learning", yet it is translated as "recharge one's batteries", which in English refers to "restoring physical strength and energy" (Jin 2008). Jin (2007) points out that such erroneous translations could be misleading or confusing, which illustrates the need for precise translation of neologisms.

When it comes to the presentation of neologisms, some lexicographers (e.g., Marello 2020: 176) have observed that dictionaries normally introduce neologisms, but neglect to represent them in a comprehensive way. For instance, dictionaries rarely give examples of usage-in-context for them, nor do they provide certain morphosyntactic items of information because lexicographers assume

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1948 (Article)

that such linguistic patterns governed by morphological and grammatical rules should be known to all. Wang and Chen (2024: 50) also note that when including neologisms, aspects such as definition, examples and outside matter also deserve the attention of lexicographers.

2.3 Criteria and strategies to include and represent neologisms

Faced with the above challenges, lexicographers have proposed various criteria for the inclusion of neologisms from the macrostructural perspective, their arrangement, presentation and strategies of translation from the microstructural perspective, as highlighted by Metcalf (2004), Xiao (2017), Yang (2014), and others. Metcalf (2004) outlines five criteria, known as the "FUDGE rule" (Frequency, Unobtrusiveness, Diversity of users and situations, Generation of meanings and forms, Endurance of concept), for including new words into a dictionary. However, Wang and Lu (2004: 404) argue that frequency is not the sole criterion for inclusion as dictionaries could not entirely omit nonce words, which users may still wish to consult. Regarding the endurance of a concept, some scholars (Chao 1992; Su and Huang 2003) believe that the inclusion of new words should be subject to the test of time to demonstrate their long-term value in social interactions. However, others argue that dictionaries should promptly include neologisms to serve users' needs. For example, Liu (1984) and Lv (1984) note that dictionaries are often overly cautious in adding new words, focusing too much on maintaining stability. They suggest that lexicographers should broaden their inclusion of neologisms.

As for the translation of neologisms, researchers such as Jin (2007), Yang (2014) and Xiao (2017) propose various rules or strategies for identifying appropriate equivalents. Yang (2014) holds that translations of political neologisms should be accurate, conventional, authoritative, unified and concise. Jin (2007) addresses the problem of potential confusion arising from the *equivalents* offered for the entries of neologisms. He argues that to lower the chance of misunderstandings, it is necessary to include usage notes or explanations in addition to the equivalents listed.

Xiao (2017) explores eight translation strategies for new words and expressions from the *Chinese–English Dictionary* (unabridged, volume 1, 2015), chief-edited by Gusun Lu, including pure literal translation, literal translation combined with explanation, literal translation combined with free translation, pure free translation, free translation combined with explanation, free translation combined with transliteration, coinage, and back translation. He believes that the exploration of these strategies holds certain guiding significance and reference value for improving the translation quality of new words and expressions, better facilitating cultural exchange between China and the West.

Relevant studies provide excellent theoretical and practical insights for research on the inclusion of neologisms in Chinese foreign language dictionaries. This study, building upon these insights, concentrates on the specific method-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1948 (Article)

ologies employed by the revision team on the inclusion of neologisms and their French equivalents in the GDCFC, with the objective of identifying more effective approaches to satisfy the user needs regarding neologisms.

3. Neologisms in the revision of the GDCFC

In the first edition of the GDCFC, the editorial team meticulously extracted neologisms from printed sources like books and newspapers, first recording them on index cards before digitizing the information. This manual data collection process was not only labor-intensive but it also delayed the dictionary's publication by nearly 16 years. In the big data era, Huang (2016) highlights the necessity of the corpus-driven approach for dictionary revision. According to Zhang and Yong (2007: 421), corpus data provides lexicographers with a comprehensive understanding of a word's characteristics, including its syntactic patterns, collocations, semantic meanings, and contexts of use. This approach, as Fontenelle (2015: 14) points out, enables lexicographers to make informed decisions about including and defining neologisms based on evidence from parallel corpora, rather than relying solely on intuition or prior knowledge. Wei (2009) further emphasizes that such a methodology ensures a more objective and scientific dictionary compilation process.

Therefore, in order to achieve more reliable results before including neologisms in the GDCFC, the team identifies neologisms through various sources such as parallel corpora, monolingual corpora, monolingual and bilingual dictionaries, and the Internet. The main source is a parallel Chinese–French corpus, which includes expressions in both languages and is primarily sourced from Foreign Language Teaching and Research Press. The data was mainly constructed by a group of Chinese lexicographers who translate new words and expressions from French news media into Chinese, and vice versa. Another part of the parallel corpus was collected by the revision team from authoritative online texts, such as government reports and important conferences, ensuring the quality of equivalents for Chinese neologisms. This approach addresses the challenge mentioned by Gao (2003), where suitable equivalents for Chinese neologisms are often lacking, necessitating lengthy definitions.

Through seven years (2016–2023) of work, the team has sorted out more than 3,000 neologisms, together with their French equivalents, appropriate examples and usages. The neologisms are collected mainly from eight key domains: **politics**, including terms like 反腐 (anti-corruption) and 多边主义 (multilateralism); **economics**, with terms such as 供给侧 (supply side) and 跨境电商 (cross-border e-commerce)); **health care**, encompassing terms like 埃博拉病毒 (Ebola virus), 新冠肺炎 (COVID-19), and 抗疫 (anti-epidemic); **environment,** with terms such as 低碳 (low carbon) and 新能源 (new energy); the **Internet and information technology**, including 网瘾 (internet addiction), 生成式人工智能 (generative AI), and 大语言模型 (large language model); **education,** featuring terms like 慕课 (MOOC) and 在线教育 (online education); **transportation**, including 共享单车 (shared bike),

无人驾驶汽车 (driverless car), and 磁悬浮列车 (maglev train); and **daily life,** with terms such as 蚁族 (antizen) and 布基尼 (burkini).

Apart from new words, the team has also noted the development of new meanings for existing terms. For instance, "跳水" originally refers to "diving", but it has acquired an additional meaning in the financial sector, where it denotes a sudden drop in prices. "应用" is originally used as a verb, meaning "to apply" or "to use", but it has now gained a noun sense, referring to "app", particularly in the context of software applications for mobile devices. "钓鱼", traditionally associated with the act of fishing, now refers to "phishing", the practice of using deceptive methods to trick individuals into revealing personal information online.

The GDCFC revision team has also included new words formed by affixes derived from nouns, such as "奴"(slave), "云"(cloud), and "零"(zero). For instance, the suffix "奴" denotes a slave in a metaphorical sense in Chinese. The team has included new words formed with this suffix, such as "房奴" (mortgage slave), "孩奴" (child slave), "卡奴" (credit card slave), and "车奴" (car slave). Furthermore, terms with the prefix "零" (zero), like "零容忍" (zero tolerance) and "零利息" (zero interest), have also been added.

4. Problems encountered when including neologisms

When including neologisms into the GDCFC, the team has encountered mainly four types of problems.

4.1 Challenge of including neologisms that are absent from authoritative Chinese dictionaries

Bilingual dictionaries typically take as reference the latest authoritative monolingual dictionaries for their inclusion of neologisms. The revision of the GDCFC is no exception. It takes as references the *Contemporary Chinese Dictionary* (7th edition, CCD) and the *Normative Dictionary of Contemporary Chinese* (4th edition, NDCC). Neologisms not found in the first edition of the GDCFC but included in the recent editions of these authoritative Chinese dictionaries, such as "大数据" (big data), "供给侧" (supply-side) and "点赞" (like), are generally added into the GDCFC unless there is no appropriate French equivalent available. However, for terms not officially recognized by these dictionaries but widely used and supported by corpus evidence or online resources, such as "获得感" (sense of achievement), "布基尼" (bikini), "新冠肺炎" (COVID-19), "伤不起" (in a humorous manner: can't afford to be hurt), and "神马" (a homophone for "什么", whose meaning is "what"), lexicographers should establish some criteria to decide their inclusion because not all of them could be included in the dictionary.

4.2 Dilemmas regarding variant neologisms and their French equivalents

Neologisms often have variant forms. For example, zika virus have three variant forms in Chinese: "兹卡病毒", "寨卡病毒", "齐卡病毒". Which form is more acceptable? What criteria should lexicographers adopt? This is a very tricky problem. And for "一带一路", there are at least three translational versions for it: *une ceinture et une route*, *l'initiative Ceinture et Route*, *la Ceinture et la Route*. The first equivalent is extracted from the parallel corpus, the second and the third ones are from the French versions of Chinese authoritative journals. Which one should be chosen as an equivalent in the dictionary? Lexicographers must set up a criterion to treat the problems of variants concerning the neologisms and their equivalents.

4.3 Dilemma of including neologisms as entries or examples

Determining whether neologisms should be recognized as separate entries, subentries, or merely included as examples within existing definitions in dictionaries is a challenge requiring careful editorial consideration. For instance, "咖啡机" (coffee machine) was included as an example of the entry "咖啡" (coffee) in the first edition of the GDCFC. While the revision team has found in the bilingual corpora and in the search engines Baidu and Bing occurrences of several types of coffee machines, such as "胶囊咖啡机" (capsule coffee machine), and "全自动咖啡机" (fully automatic coffee machine). Therefore, users may need to consult these expressions and their French equivalents. Lexicographers have to decide whether "咖啡机" continues to be cited as an example under the broader category of "咖啡", with its various types also listed as examples, or whether "咖啡机" itself should be treated as a separate entry, with its kinds then becoming examples under "咖啡机".

Another example is "酸辣汤" (sour and spicy soup), which was treated as a separate entry in the first edition. However, with the appearance of similar entries as "酸辣粉" (sour and spicy rice vermicelli), "酸辣面" (sour and spicy noodles) and "酸辣酱" (sour and spicy sauce), lexicographers have to consider if all these neologisms should be treated as entries. If all similar terms are treated as separate entries, the dictionary could become overly extensive with unwieldy number of entries, leading to a lack of diversity in entry types and a failure to effectively illustrate the interconnections of entries related to similar concept.

4.4 Challenge of capturing relations between neologisms and existing entries

Many bilingual dictionaries focus on offering equivalents for neologisms, neglecting relations between these words and existing entries. The revision team endeavors to reveal such connections by making use of the parallel corpus. The problem is that the size of the corpus from which neologisms are selected is limited, which sometimes make it difficult to fully demonstrate the relations centered by one entry or by one concept in some semantic fields.

For example, for the entry "充电" (charge or recharge), in the first edition, besides the French equivalents, only four examples (充电器, 给蓄电池充电, 重新充电, 充电设备) were offered. New words and phrases relevant to "充电" (such as 充电桩, 充电宝, 充电站, 无线充电, 隔空充电, and 快速充电) should also be represented as examples or relevant entries in the dictionary, which also offers a quick access for users. Nevertheless, in the parallel corpus only "充电宝" and "充电桩" can be found, others (such as 无线充电, 隔空充电, and 快速充电) not. A search conducted in the monolingual Chinese web corpus, as well as in the search engines Baidu and Bing, has revealed that the latter expressions appear in large quantity in these sources. If the team only includes the data in the parallel corpus, the complete wordnet representing the concept "充电" would not be formed. More sources should be explored.

5. Criteria proposed to include neologisms

To better solve the afore-mentioned issues and challenges encountered during the revision work, the GDCFC revision team, based on their observations, continuous practice and discussions, together with previous lexicographers' research insights, and especially with the assistance of multiple resources, established four criteria to comprehensively include and present neologisms in the second edition of the GDCFC in an optimized way to better cater for the needs of the users.

5.1 Criterion of descriptivism complemented with prescriptivism

The GDCFC adopts the criterion of combining descriptivism and prescriptivism. The GDCFC revision team primarily adopts a descriptive criterion for the inclusion of new words, which means to describe the language use in an objective way by trying to include as many neologisms as possible based on their frequency. At the same time, the team strictly adheres to linguistic and terminological norms. The majority of new terms and their translations are sourced from parallel corpora, primarily from authoritative websites, and the examples of entries are mainly extracted from the official bilingual documents (see the following two entries). This guarantees that the inclusion of neologisms and their translations follow strictly official and academic standards.

【供给侧】du côté de l’offre ▹深入推进供给侧结构性改革 poursuivre en profondeur la réforme structurale du côté de l’offre

【获得感】 sentiment *m* de satisfaction, sentiment *m* d’obtention ▹让人民群众有更多“获得感” Que le peuple puisse avoir un plus grand sentiment de satisfaction.

As for the selection of word forms, the team adheres to national or international standards. For example, there are three variant forms of Chinese loanwords from zika virus: "兹卡病毒", "寨卡病毒" or "齐卡病毒". Only "寨卡病毒" is included in the terminology bank created by the National Committee for the Standardization of Scientific and Technical Terminology, which officially shows its standard for the right form of *zika* in Chinese. Therefore, "寨卡病毒" is selected as the entry instead of the other two variant forms. And for the three equivalents of "一带一路" mentioned in 4.2, the team has carefully identified the instances when equivalent terms are presented and has observed an evolution of translation strategies for this term. The latest translations from authoritative Chinese websites have been ultimately adopted: "la Ceinture et la Route" and "l'initiative Ceinture et Route", rejecting the earlier translation "une ceinture et une route".

Meanwhile, in order to standardize the Internet language, neologisms that are often used for banter, jokes, or by onomatopoeia, such as "神马" (a homophone for "什么", whose meaning is "what") and "伤不起" (used in a humorous manner: can't afford to be hurt), are generally not included in the revised GDCFC.

5.2 Criterion of Frequency

Corpus-based frequency analysis is a common practice in lexicography. The revision team uses the criterion of frequency to select and define new entries. It focuses on three aspects: (1) choosing among variant forms of a neologism or of its equivalents, (2) deciding whether to include a neologism as an entry or an example, and (3) whether to add new meanings. By verifying these issues through multiple sources, especially search engines and corpora (both monolingual and bilingual), the team has managed to ensure the selection of neologisms and the precision of translations, thereby preventing omissions or inaccuracies that can arise from over-reliance on lexicographers' introspection or on the content of the print dictionaries.

5.2.1 Choice among variant forms

Lexicographers working on bilingual dictionaries regularly face the challenge of deciding on the right form among the variant forms of neologisms that have not yet been documented in established Chinese dictionaries. The occurrence of variant forms in large-scale corpora can assist lexicographers in their decision. The revision team primarily relies on Chinese web corpus of the corpus tool Word Sketch Engine[1] to verify the frequency of variant forms.

For example, the word *brunch* is translated as "早中餐" in the parallel corpus. However, a search by the team in Chinese web corpus reveals other translations such as "早午餐", "早中饭" and "早午饭". None of the four variants is included in the CCD or in the NDCC. The frequency of "早午餐" in the corpus is significantly higher than the other three variants (see Table 1). Consequently, the revision team has opted to include "早午餐" as the entry in the dictionary, while excluding the other three translations.

Table 1: Search results from the Chinese web corpus of the variant forms for the Chinese translation of *brunch*

variant form of the neologism	frequency of the variant form in Chinese web corpus (hits)
早午餐	3059
早中餐	5
早午饭	19
早中饭	2

5.2.2 Decision of inclusion of neologisms as entries or examples

Frequency serves as the basis for lexicographers to decide whether new words should be included as examples or as entries. For example, 【酸辣汤】 mentioned in 4.3 was treated as a separate entry in the first edition. While in the Chinese web corpus, new words similar to "酸辣汤" emerge, like "酸辣粉", "酸辣面", and "酸辣酱". The frequency of "酸辣粉" (8092 hits) is considerably higher than "酸辣汤" (1809 hits). Therefore, the revision team has created the entry 【酸辣】, with two most frequently used words "酸辣汤" and "酸辣粉" presented as examples (see the following entries). This approach helps users better understand these terms and the semantic relations between them.

【酸辣汤】 potage acide et âcre ; potage poivré et vinaigré

【酸辣】 acide et âcre *a*; poivré(e) et vinaigré(e) *a* ▹酸辣粉 vermicelles [nouilles de riz] poivrés et vinaigrés ‖ 酸辣汤 potage acide et âcre ; potage poivré et vinaigré

Another example mentioned in 4.3 is "咖啡机", which was presented as an example under the entry 【咖啡】 (coffee). If the high-frequency terms (see 4.3) related to "咖啡" including "胶囊咖啡机" and "全自动咖啡机" are all presented together with "咖啡机" as examples under the entry "咖啡", it would result in an excessive number of examples and unwieldy amount of information for the already example-rich entry of "咖啡", making it difficult for users to find the information they need. Therefore, based on the criterion of frequency, the revision team treats "咖啡机" as a separate entry rather than an example of 【咖啡】, and the related derivative terms are presented as examples under this new entry.

【咖啡】➊caféier *m* ; café *m* ▹~种植 la culture du caféier ‖ ~园 plantation de caféiers ➋café *m* en poudre ; café *m* ➌café (boisson) *m* ▹一杯~ une tasse de café ‖ 牛奶~ café au lait ‖ 速溶~ café instantané ‖ 清~ café noir ‖ ~杯 tasse à café ‖ ~伴侣 succédané de lait pour le café◆~厅[馆］ café *m* ‖ ~壶 cafetière *f* ‖ ~机 cafetière *f*

【咖啡机】 cafetière *f*; machine *f* à café ▹研磨咖啡机 machine à moudre le café ‖ 胶囊咖啡机 machine à café à capsule ‖ 全/半自动咖啡机 machine à café entièrement automatique/semi-automatique

5.2.3 Determining whether to add new meanings

The revision team uses the criterion of frequency to decide whether new meanings of old words, such as "应用" (app) and "跳水" (sudden drop in prices), should be considered for inclusion in the dictionary. For instance, the term "应用", besides its usages as a verb, has a new meaning used as a noun: "app", which is widely used and supported by corpus data and search engine results. Therefore, the team has decided to include this new meaning in the revised GDCFC.

5.3 Supplementary criterion

Traditional bilingual dictionaries often provide equivalents for entries without any explanatory notes. This approach sometimes falls short in clarifying the meaning or usage of an entry. Adamska-Sałaciak (2015) advocates for "supplementary meaning-elucidating strategies" in the compilation of bilingual dictionaries. She contends that when the provision of equivalents alone does not suffice to convey a source language (SL) meaning, lexicographers should enhance the entries with additional semantic information. Similarly, Lu (2003) emphasizes the necessity of supplementary explanations for neologisms in cultural domains and loanwords of foreign origins. Based on these insights and the revision practices, the GDCFC revision team has set up the supplementary criteria primarily in mainly two aspects: adding explanations to the equivalents of neologisms, especially for new loanwords or cultural words, and offering full forms for new abbreviated words. This approach aims at a deeper understanding of the neologisms.

5.3.1 Adding explanations to the equivalents of neologisms, especially for new loanwords or cultural words

The team has added explanations to the equivalents for new words or expressions, especially for new loanwords or cultural words to help understanding. For example, when including "布基尼" into the GDCFC, lexicographers would not only offer *burkini* extracted from the parallel corpora as its equivalent, because it would not help much for the understanding of its meaning. To make it much clearer, an explanation with brackets is added: burkini *m* (tenue de bain des femmes musulmanes).

5.3.2 Offering full forms for new abbreviated words

The team has offered full forms for new words created through abbreviations to enhance comprehension of their meanings. For example, the full form 农业学校 is added in brackets for its abbreviated form 【农校】, and 亚洲基础设施投资银行 for【亚投行】.

【农校】[〈abrév.〉pour 农业学校] école *f* agricole [d'agriculture]
【亚投行】[〈abrév.〉pour 亚洲基础设施投资银行] Banque *f* Asiatique d'Investissement pour les Infrastructures (BAII ou AIIB en anglais)

5.4 Criterion of relevance

As Li and Huang (2017) claim, words do not exist in isolation within a linguistic system. There are various associations between lexical units and concepts, which together form an intricate network of semantic relationships. Therefore, to comprehensively describe a neologism, lexicographers must highlight its systematic semantic associations. From this perspective, the team has focused not just on providing equivalents to describe a neologism, but has also applied the criteria of relevance in their work. To better organize the neologisms and their related entries and present them within the context of the semantic framework, the team tries to provide semantic field components and relevant collocations associated with the headword. The aim is to help dictionary users build a semantic network of mental vocabulary, creating a net that covers the entire semantic field.

For instance, the entry "充电" mentioned in 4.5 already provides some examples for its original meaning in the first edition of the GDCFC and in the parallel corpus, while some new words concerning the concept "充电" and other key elements are not included. The revision team, based on monolingual corpus data and online sources, tries to reveal all these elements within the semantic framework linking to the concept. Therefore, new words concerning the charging devices or equipment, such as "充电站" (charging station), "充电桩" (charging pile), "充电线" (charging cable), and "充电宝" (power bank), are organized as subentries under the entry "充电". In addition, the VN collocation construction referring to charging objects, such as "给手机、手提电脑、电动汽车等充电" (to charge a cellphone, a portable computer, an electric car, etc.) and the ADJ+N construction referring to various types of charging, including "无线充电" (wireless charging), "有线充电" (wired charging), "快速充电" (fast charging), "移动充电" (mobile charging) and "常规[慢速]充电" (slow charging), are treated as examples of the entry. It should be noted that "充电器" was treated as an example of "充电" in the first edition, while in the second edition, it is regarded as an entry with "手机充电器" (cellphone charger) and "无线充电器" (wireless charger) as examples.

The ameliorated entries are presented below.

【充电】1. charger [recharger] *v* ; charge [recharge] *f* ▷给手机、手提电脑、电动汽车、蓄电池等充电 charger [recharger] un téléphone portable, un ordinateur portable, un véhicule électrique, une batterie d'accumulateurs ‖ 快速充电 recharge rapide ‖ 常规[慢速]充电 recharge lente ‖ 无线充电 recharge sans fil ‖ 有线充电 recharge avec fil 2.<fig.> enrichir ses connaissances

【充电器】chargeur *m* ▹手机充电器 chargeur pour téléphone portable ‖ 无线充电器 chargeur sans fil
【充电线】câble *m* de recharge
【充电宝】batterie *f* portable, batterie *f* externe ▹乘坐飞机时，充电宝必须随身携带。La batterie portable n'est autorisée qu'en cabine lors d'un vol.
【充电站】centre *m* de recharge ▹建立充电站 mettre en place [installer] des centres de recharge
【充电桩】borne *f* de recharge ▹电动车充电桩 borne de recharge pour véhicule électrique ‖ 快速充电桩 point de recharge rapide

As for concept-related word groups where there are differences in French equivalent expressions, the revision team provides a more diverse range of examples. For instance, in Chinese, "无人" refers to phenomena that do not require human service or operation, while in French, to express the concept, different translation equivalents must be used depending on the context. For example, "无人超市" (unmanned supermarket) can be translated as "supermarché sans caisse" or "supermarché libre-service"; "无人汽车" (driverless car) can be translated as "voiture autonome" or "voiture sans conducteur"; and "无人机" (unmanned aerial vehicle) is translated as "drone". All these expressions with different equivalents are presented as examples for the entry 【无人】 to serve as translation references for users, assuring the accuracy of each neologism and help avoid translation errors.

Through this approach, dictionary users are able not only to understand the meaning, usage and the accurate translations of new words but also to explore their connections with existing entries, centered around a key concept.

As Huang (2016: 5) has pointed out, for lexicographers, "writing a few articles, using only a limited number of entries to showcase their semantic relationships or semantic networks, seems to be manageable [...] However, facing the vast ocean of vocabulary, it is far from easy for them to exhaustively describe the interconnections among all these terms". Therefore, the criterion of relevance is only experimental, and needs ongoing enhancement, revision, and refinement in the future.

6. Conclusion

This study describes the special practices and experiences of the revision team of the GDCFC on the inclusion of neologisms for its second edition using a wide spectrum of sources, ranging from the latest monolingual and bilingual dictionaries, monolingual and bilingual corpora, to online resources. To respond to the challenges encountered by the team, we propose four criteria for the inclusion of neologisms in the GDCFC: descriptivism plus prescriptivism, frequency, supplementation, and relevance. The application of these criteria helps achieve a more objective, more precise inclusion and a more comprehensive description on the neologisms in bilingual dictionaries, forming a complete syntactic and

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1948 (Article)

semantic network centering on the neologism or the core concept. It also provides a better reflection of the dynamic changing process of Chinese and French languages while guiding users on the use and learning of neologisms. The recommendations offered herein are expected to provide valuable insights into future lexicographic work on neologisms in bilingual dictionaries.

Acknowledgements

This study was supported by Chinese Academic Translation Project of the National Social Science Foundation (22WYYB009). We are grateful to Prof. Dion Nkomo and two anonymous reviewers for their insightful comments and suggestions, as well as to the typesetters, Ms Tanja Harteveld and Ms Hermien van der Westhuizen.

Endnote

1. https://www.sketchengine.eu/

References

A. Dictionaries

Chinese Academy of Social Sciences (Ed.). 2016. *Contemporary Chinese Dictionary*. Seventh edition. Beijing: The Commercial Press. (CCD)

Huang, J.H. (Ed.). 2014. *Grand Dictionnaire Chinois–Français Contemporain*. Beijing: Foreign Language Teaching and Research Press. (GDCFC)

Li, X.J. (Ed.). 2022. *Normative Dictionary of Contemporary Chinese*. Fourth edition. Beijing: Foreign Language Teaching and Research Press, and Language and Culture Press. (NDCC)

Lu, G.S. (Ed.). 2015. *Chinese–English Dictionary* (unabridged). Shanghai: Fudan University Press.

B. Other literature

Adamska-Sałaciak, A. 2015. Explaining Meaning in Bilingual Dictionaries. Durkin, P. (Ed.). 2015. *The Oxford Handbook of Lexicography*: 144-160. Oxford: Oxford University Press.

Balteiro, I. 2011. Prescriptivism and Descriptivism in the Treatment of Anglicisms in a Series of Bilingual Spanish–English Dictionaries. *International Journal of Lexicography* 24(3): 277-305.

Cao, D.M. 2021. Spread Chinese Culture, Relay Sino–French Exchanges — Congratulations on the Official Publication of the *Grand Dictionnaire Chinois–Français Contemporain*. Tian, B. and F. Huang (Eds.). 2021. *Examples of Inheritance and Innovation in the Compilation of Chinese Foreign Language Dictionaries in the New Era*: *A Collection of Research Papers on the* Grand Dictionnaire Chinois–Français Contemporain*:* 234-237. Beijing: Foreign Language Teaching and Research Press.

Chao, J.Z. 1992. On the Characteristics of Normative Dictionaries — With a Discussion on the Word Inclusion Principles of the *Contemporary Chinese Dictionary*. *Lexicographical Studies* 5: 145-151.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1948 (Article)

Chen, B. 1980. A Dictionary Grouping English Neologisms — 6000 Words, Supplement to *Webster's Third New International Dictionary*. *Lexicographical Studies* 4: 137-143.

Cheng, R. 2006. Several Aspects of the 5th Edition of the *Contemporary Chinese Dictionary* in Reinforcing Standardization. *Lexicographical Studies* 1: 11-17.

Du, K.H. 2019. On the Translation of New Words and Meanings in Chinese English Dictionaries — An Empirical Study Based on Five Dictionaries. *Fudan Forum on Foreign Language and Literature* 2: 141-145.

Fontenelle, T. 2015. Bilingual Dictionaries, History and Development: Current Issues. Durkin, P. (Ed.). 2015. *The Oxford Handbook of Lexicography*: 44-61. Oxford: Oxford University Press.

Gao, Y.W. 2003. On the Two Problems in the Translation of Chinese New Words. *Shanghai Journal of Translators for Science and Technology* 2: 45-47.

Gouws, R.H. and L. Potgieter. 2010. Does Johnson's Prescriptive Approach Still Have a Role to Play in Modern-Day Dictionaries? *Lexikos* 20: 234-247.

Hanks, P. 2013. Lexicography from Earliest Times to the Present. Allan, Keith. (Ed.). 2013. *The Oxford Handbook of the History of Linguistics*: 503-536. Oxford: Oxford University Press.

Huang, J.H. 2016. Some Thoughts on Revision of the Grand Chinese–French Dictionary. *Lexicographical Studies* 5: 1-6.

Huang, J. and H. Xu. 2019. Reflections on the Making of the *Grand Dictionnaire Chinois–Français Contemporain*. *Lexikos* 29: 324-338.

Jin, Q.B. 2007. Analysis of the Merits and Deficiencies on the Representation of Neologisms in *Longman Dictionary of Contemporary English*. *Lexicographical Studies* 1: 96-104.

Jin, Q.B. 2008. General English Translation Theories: A New Approach to Translation of New Words and Expressions in Chinese into English. *Terminology Standardization and Information Technology* 2: 16-21.

Jin, Q.B. 2009. Coverage and Translation of New Words in Chinese in *A New Century Chinese–English Dictionary*. *Journal of Shenzhen Polytechnic* 6: 74-80.

Li, L. and J.H. Huang. 2017. On the Construction of Semantic Network in a Chinese–French Dictionary. *Lexicographical Studies* 2: 37-41.

Liu, X.J. 1984. New Words and Meanings and the Inclusion of Words in Language Dictionaries. *Lexicographical Studies* 6: 62-67.

Lu, J.Q. 2003. Comparative Analysis on Dictionaries of New Words: Similarities and Differences Between Monolingual and Bilingual Dictionaries. Zeng, D.J. (Ed.). 2003. *Bilingual Lexicography Research: Selected Papers of the 5th National LSC Symposium on Bilingual Lexicography*: 61-67. Shanghai: Shanghai Foreign Language Education Press.

Lv, S.X. 1984. Attention Should be Paid to New Words and Meanings. *Lexicographical Studies* 1: 8-14.

Marello, C. 2020. New Words and New Forms of Linguistic Purism in the 21st Century: The Italian Debate. *International Journal of Lexicography* 33(2): 168-186.

Metcalf, A. 2004. *Predicting New Words: The Secrets of Their Success.* Boston: Houghton Mifflin Harcourt.

Mugglestone, L. 2015. Description and Prescription in Dictionaries. Durkin, P. (Ed.). 2015. *The Oxford Handbook of Lexicography*: 546-560. Oxford: Oxford University Press.

Su, X.C. and Q. Huang. 2003. The Maturity and Standardization of New Words and the Selection Standard in Dictionaries — On the "Appendix of New Words" in *Contemporary Chinese Dictionary* (2002 supplement). *Lexicographical Studies* 3: 106-113.

Ten Hacken, P. 2020. Norms, New Words, and Empirical Reality. *International Journal of Lexicography* 33(2): 135-149.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1948 (Article)

Walter, H. 2016. La Norme Linguistique dans le Dictionnaire de l'Académie Française. *La Linguistique* 52(1): 55-67.

Wang, A. and X. Chen. 2024. On the Inclusion of Neologisms in *Oxford Advanced Learner's Dictionary* (10th edition). *Lexikos* 34: 41-50.

Wang, F.F. and G.S. Lu. 2004. How Do Dictionary Compilers Deal with Nonce Words in Speech Flow? *Foreign Language Teaching Abroad* 4: 41-47.

Wang, F.F. and G.S. Lu. 2007. Inheritance Plus Innovation. On the Revision of *The English–Chinese Dictionary*. *International Journal of Lexicography* 20(1): 1-38.

Wang, T.K. 1992. Criteria for Determining New Words and Principle of Compiling Dictionary of New Words. *Applied Linguistics* 4: 14-20.

Wang, Y.Y. 2010. A Corpus-based Approach to Definition in the Chinese Dictionary. *Lexicographical Studies* 1: 111-118.

Xiao, F.S. 2017. Strategies Regarding the Translation of Neologisms in a Chinese–English Dictionary. *Shanghai Journal of Translators* 2: 79-85.

Yang, M.X. 2014. Translation Principles and Strategies for Chinese Diplomatic Neologisms. *Chinese Translation* 3: 103-107.

Yu, G.Y, T.K. Wang and S.X. Sun. 2003. The Basic Principles of the Standardization of New Words and Expressions. *Applied Linguistics* 1: 89-95.

Zhang, Y.H. and H.M. Yong. 2007. *Contemporary Lexicography.* Beijing: Commercial Press.

Zhao, G. 2014. Review of Pan Shaozhong (ed.) 2014. *New Age Chinese–English Dictionary* (2nd edition). *International Journal of Lexicography* 27(4): 435-451.

Zhao, G. 2015. Making a User-friendly Bilingual Dictionary for Chinese Translators: On the Revision of *A New Century Chinese–English Dictionary*. *International Journal of Lexicography* 29(4): 452-489.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1949 (Article)

Bridging across Polysemic Senses in Bilingual Specialized Dictionaries for ESP Learners

Huaguo Lu, *School of Foreign Studies, Nanjing University of Science and Technology, China (louisluhuaguo@163.com) (https://orcid.org/0009-0007-9825-3661) (Corresponding Author)*
and
Yundong Geng, *Huazhong Agricultural University, Wuhan, China (benedict_geng@mail.hzau.edu.cn) (https://orcid.org/0000-0002-8519-0821)*

Abstract: Research has shown that links between polysemic senses (sense links) can and should be used to facilitate the acquisition of polysemy. However, sense links have received little attention in specialized lexicography because the concern about domain specificity has considerably reduced the number of polysemic senses that can be entered in specialized dictionaries. The descriptive shift in terminology research and the implications of cognitive semantics for learner's dictionaries have paved the way for dealing with sense links further in specialized dictionaries for learners (SDLs). Using computing-related lexical items as examples, this article proposes three guidelines for treating polysemy in SDLs with the aim of entering polysemic senses that do not belong to a given subject field while maintaining the focus on the subject field. It also presents four models for describing sense links in bilingual specialized dictionaries for ESP learners (BSDLs). Depending on the magnitude of overlap between the target language (TL) equivalents of the source and target senses as well as the effects of other factors, sense links are represented by ordering senses logically, appending the source sense[1], combining logical ordering with a short explanation, or providing both the source sense and a short explanation. The guidelines and models can help address the major situations that lexicographers encounter when describing sense links in BSDLs and hopefully contribute to learners' acquisition of technical senses.

Keywords: SENSE LINKS, ACQUISITION OF TECHNICAL SENSES, BILINGUAL SPECIALIZED DICTIONARIES FOR LEARNERS, DOMAIN SPECIFICITY, GUIDELINES, MODELS, SEMANTIC DISTANCE, OVERLAP OF TARGET LANGUAGE EQUIVALENTS

Opsomming: Die oorbrugging van polisemiese betekenisse in tweetalige gespesialiseerde woordeboeke vir ESD-leerders. Navorsing het getoon dat polisemiese betekenisse (betekenisskakels) gebruik kan en moet word om die aanleer van polisemie te vergemaklik. Betekenisskakels het egter min aandag in gespesialiseerde leksikografie ontvang aangesien die fokus op domeinspesifiekheid die aantal polisemiese betekenisse wat in gespesialiseerde woordeboeke opgeneem kan word, aansienlik verminder het. Die deskriptiewe skuif in terminologienavorsing en die implikasies wat die kognitiewe semantiek vir aanleerderwoordeboeke inhou, het die weg gebaan vir die verdere hantering van betekenisskakels in gespesialiseerde woordeboeke vir leerders (GWL's). Deur rekenaarverwante leksikale items as voorbeelde te gebruik, word

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1949 (Article)

daar in hierdie artikel drie riglyne vir die hantering van polisemie in GWL's voorgestel met die doel om polisemiese betekenisse wat nie tot 'n gegewe vakgebied behoort nie, op te neem, terwyl die fokus steeds op die vakgebied bly. Vier modelle vir die beskrywing van betekenisskakels in tweetalige gespesialiseerde woordeboeke vir ESD-leerders (TGWL's) word ook voorgestel. Afhangende van die omvang van die oorvleueling tussen die doeltaalekwivalente (DT-ekwivalente) van die bron- en doelbetekenisse sowel as die gevolge van ander faktore, word die betekenisskakels voorgestel deur die betekenisse logies te orden, deur die bronbetekenis[1] by te voeg, deur logiese ordening met 'n kort verklaring te kombineer, of deur die bronbetekenis en 'n kort verklaring te verskaf. Die riglyne en modelle kan help om die vernaamste situasies wat leksikograwe teëkom wanneer hulle betekenisskakels in tweetalige gespesialiseerde woordeboeke vir ESD-leerders beskryf, te hanteer, en hopelik bydra tot leerders se aanleer van tegniese betekenisse.

Sleutelwoorde: BETEKENISSKAKELS, AANLEER VAN TEGNIESE BETEKENISSE, TWEETALIGE GESPESIALISEERDE AANLEERDERWOORDEBOEKE, DOMEINSPESIFIEKHEID, RIGLYNE, MODELLE, SEMANTIESE AFSTAND, OORVLEUELING VAN DOELTAALEKWIVALENTE

1. Introduction

Polysemy has been recognized as a significant problem in learning vocabulary in both General English (GE) and English for Specific Purposes (ESP) (Mićović and Beko 2022: 125). It is acknowledged that "there is a great deal more involved in knowing a word in an L2 than being able to match it with an L2 synonym or provide an L1 translation equivalent" (Read 2004: 211). One suggestion for increasing the depth of EFL learners' vocabulary knowledge is to acquire the multiple meanings of polysemous lexical items (Richards 1976; Nation 1990; Li and Kirby 2015). However, achieving this goal is not easy.

Compared to native speakers' use of polysemy, EFL learners' difficulty in acquiring polysemy can partly be attributed to their limited ability to deduce peripheral or infrequent senses from central or frequent ones (Miao 2015: 221). Therefore, teachers are recommended to "help learners to get accustomed to the idea that different uses of words may have a shared underlying meaning" (Nation 2013: 306) and "to see how the technical sense of the words relates to the core meaning of the word" (Chung and Nation 2003: 113). Empirical studies have shown that awareness of the central or core senses of polysemes contributes to the acquisition of the peripheral or non-core senses (Verspoor and Lowie 2003; Maby 2016). Cognitive linguistics provides further insights into this issue. Various mechanisms such as metaphor, metonymy, specialization, generalization, profile shift and image-schema transformations have been proposed to explain how polysemic senses are interrelated (Lakoff 1987; Radden and Kövecses 1999; Taylor 2003; Tyler and Evans 2004; Gries 2015: 474). These insights can be used to train EFL learners to see the connections between polysemic senses. In fact, some researchers have experimented with the cognitive semantic view of polysemy in EFL settings, and the majority of these studies have confirmed that explaining the motivations for semantic extensions promotes L2 learners' acquisition of polysemes (Csábi 2004; Morimoto and Loewen 2007; Beréndi et al. 2008;

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1949 (Article)

Tyler et al. 2011; and Zhao et al. 2018).

However, specialized lexicography has shown limited interest in these theoretical insights and empirical findings. There have been few discussions about the links between polysemic senses in specialized dictionaries for learners (SDLs). The first noteworthy study is Van der Meer (2010: 139), who suggested that definitions can be written "using a vocabulary (e.g. collocations) that at least strongly hints at the field of discourse from which the metaphor was originally taken" for technical senses that are extended from basic ones through transparent metaphors. Van der Meer also pointed out that "the more farfetched fanciful or complicated cases will have to remain unexplained" to avoid "changing the dictionary's ESP character" (ibid). L'Homme (2020a) is another study of particular relevance. While describing how to make meaning distinctions by presenting lexical functions, labeling argument structures or relating to semantic frames, the researcher admitted that it is difficult to account for connections between remotely linked senses using lexicographical devices such as hierarchical alphanumeric systems and cohesiveness between definitions, as terminologists usually deal with domain-specific meanings only. As revealed by both studies, links between senses remain to be further explored, and the focus on domain-specificity seems to hinder efforts to capitalize on these links more extensively.

On the one hand, it is important for ESP students to learn technical senses because they are part of key code words essential for their communication within academic discourse communities (Swales 1990). On the other hand, technical senses are difficult to learn as they are usually more peripheral and less frequent. Considering the importance and difficulty of learning technical senses, it is necessary to help ESP learners of technical vocabulary understand how a technical sense of a polyseme is extended from its other senses, regardless of whether these senses are domain-specific. This article aims to explore how bilingual specialized dictionaries for ESP learners (BSDLs) can deal with the restrictions arising from domain specificity and exploit various sense links in a user-friendly way. The rest of the article is structured as follows. Section 2 will describe the impact of the paradigm shift from Prescriptive Terminology to Descriptive Terminology on the view of polysemy. Section 3 will review the practice and proposals regarding the representation of sense links in general language dictionaries (GDs). Section 4 will explain why SDLs need to focus on one single subject field and how it is still possible to represent links between polysemic senses. Section 5 will describe the different ways of representing sense links in BSDLs, using various lexicographical devices and taking into account the semantic distance between the involved senses in a bilingual setting. The last section will summarize the research findings and limitations.

2. The influence of Descriptive Terminology on polysemy

Traditional Terminology (as a discipline) considers language in terms of its naming capacity only (Temmerman 1997: 54). It is the vocabulary that assumes the role of naming specialized concepts. However, the vocabulary used in specialized

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1949 (Article)

communication, i.e. terminology, does not seem to be very different from that used in general situations (Cabré 1999: 81), which is often described as too ambiguous. In order to promote effective and efficient communication, terminology must be standardized before it is suitable for naming concepts univocally. For instance, deliberate, albeit not always successful, attempts have been made to reduce polysemy, a common language phenomenon where one term designates more than one concept. There is now a growing consensus that "standardisation is only one aspect of what should be the concern of the theory of Terminology" (Temmerman 2000: 220). Some basic tenets of Traditional Terminology have been challenged by a descriptive paradigm. We will discuss two aspects of this paradigm that are most relevant to the present study.

2.1 Polysemy and the semasiological approach to Terminology

According to Wüster (1991: 1), Terminology begins with concepts. Only after concepts are clearly delineated within a conceptual system will terms be assigned as ideal linguistic labels. Prioritizing concepts in terminology work entitles terminologists to select or create labels for concepts, leading to an overwhelming dominance of nouns over other parts of speech in terminological resources (Rey 1979; Sager 1990: 51). Moreover, Traditional Terminology holds that the relationship between concepts and terms can be manipulated through standardization to achieve univocity. For example, when concepts across domains are designated by one term, polysemy is treated as homonymy despite the perceptible relatedness between the meanings. When concepts within a single domain share one label, polysemy is eliminated by creating new names to distinguish the concepts.

However, the onomasiological approach is rarely adopted in the practice of terminography (Sager 1990: 56; Cabré 1999: 108; Temmerman 2000: 230; L'Homme 2005: 1117) because "quite obviously, the concept is not accessible unless via the designations" and "it is the designation that serves as a starting point" (Costa 2013: 32). The alternative is the semasiological approach, which allows terminologists to identify terms in texts and work towards their meanings or the concepts they designate. It is difficult for terminologists to ignore polysemy, where one term designates several concepts. The semasiological approach also extends terms to other word classes (see L'Homme 1998). Corpus data confirm that nouns are not the only category of designations for concepts. Verbs and adjectives are more typical linguistic expressions of ACTIVITY and ATTRIBUTE, laying the foundation for defining the corresponding noun forms (L'Homme 2015: 79). Therefore, it is theoretically possible to study polysemy associated with two or more parts of speech. In addition, terms, according to the new approach, do not seem to be very different from words when considered from the formal or semantic point of view (Cabré 1999: 81). Theories of lexical semantics are also applied to researching polysemy in Terminology. In particular, insights about regular polysemy, alterations, and micro senses are

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1949 (Article)

used to identify polysemy in Terminology by making finer-grained distinctions between the multiple meanings of a term within a single field (L'Homme 2020a; L'Homme 2024).

2.2 Polysemy and the diachronic perspective on Terminology

Traditional Terminology does not study language development and language evolution. The logic behind this principle of synchronicity is that "the present meanings of terms are important", and, in order to delineate the meanings of terms, "the system of concepts is what matters in language" (Felber 1984: 98). This is not difficult to understand since "normally, when one studies terms, it is useful to view them as highly 'fixed' entities, marking clearly delineated conceptual spaces within a given domain of expertise" (Meyer and Mackintosh 2000: 111). The synchronic perspective, however, has further marginalized polysemy in Terminology.

According to Blank (2003: 268), "the best-known type of polysemy is metaphoric polysemy which derives in most cases from metaphor as a diachronic process". After all, "understanding is never a static situation but a constantly changing process in time …", so "there is a constant development in what a term can be used to refer to" (Temmerman 2000: 149-150). Therefore, it would be impossible to do justice to polysemy in Terminology without considering the diachronic dimension of the specialized language. Thanks to the descriptive shift in Terminology, researchers have come to realize that it is wishful thinking to try to fix concepts. Meyer and Mackintosh (2000) provide a textbook example of this dynamic process. Their case study of "virtual" illustrates how a general language sense (i.e. almost) evolved into a technical sense in computing (as in "virtual reality") through terminologization, which in turn was diluted to give rise to a new general language sense (as in "virtual cheesecake") or a new sense loosely related to computing (as in "virtual tours") by means of de-terminologization. It is interesting to note that the derived technical sense (as in "virtual reality"), when used in economics, was re-terminologized into a new technical sense (as in "virtual currency"). Temmerman (2000: 141, 143) traced the history of cloning, revealing that the meaning extensions of "clone" are a diachronic process of polysemization influenced by critical advancements in biology. She also investigated the metaphorical models (e.g. DNA IS A LANGUAGE) behind the process whereby a word (e.g. translate) is borrowed from one domain (e.g. language) by another domain (e.g. biology), resulting in a new sense (e.g. "decipher genetic instructions for making protein") (ibid: 184).

3. Representation of sense links in general language dictionaries

In the 19th century, lexicography, as well as etymology and semantics, "were engaged in discovering the connections between the meanings of polysemous words" (Nerlich and Clarke 1997: 351). The connection between polysemic senses

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1949 (Article)

is still a central topic in the treatment of polysemy in dictionaries. Lexicographers have been trying to represent sense links in the following two ways: (1) by laying out senses to reflect the semantic structure; and (2) by explicating how one polysemic sense extends to another. Depending on whether the lexicographical representation is informed by cognitive linguistics (CL), we will identify two periods, namely the traditional period and the CL-informed period, and review what lexicographers have achieved in the two periods respectively.

3.1 Traditional representation of sense links

The first use of the term *polysemous* in a linguistic sense can be attributed to the literary theorist August Wilhelm Schlegel (Nerlich and Clarke 1997: 351). Schlegel (1832: 42, quoted in Nerlich and Clarke 1997: 356) wrote that, when dealing with polysemous terms, lexicographers should observe the affinity between meanings and retrace the gradual and graded pathway that leads from one to the other. The author pointed out that "sometimes a single series is not enough: we have to come back several times to the common stem, so as to be able to retrace the divergent ramifications" (ibid). The processes of semantic development were referred to by Darmesteter (1886: 76) as radiation (i.e. a word accumulates meanings around a core) or concatenation (i.e. a word develops a polysemic chain of meanings). According to him, radiation and concatenation are generally mixed and combined, resulting in far more complex forms.

Lexicographers explore how the structural complexity of polysemy can be represented in dictionaries. For instance, Mel'čuk and Polguère (1995: 162-171) developed methodologies to order senses and indicate semantic distances between them in their dictionary project. Specifically, polysemic senses are ordered by considering factors such as the inclusion relationship between senses, their semantic proximity to the basic sense, the nature and regularity of semantic extension, and the underlying component for metaphorical extension. In addition, the semantic distances between senses are classified as large, medium or small according to the common part of their definitions as well as the semantic distinction between them. Finally, senses are hierarchically arranged into three layers, labelled with Roman numerals, Arabic numerals and lowercase letters to indicate the semantic distances across layers or within each layer. Mel'čuk and Polguère (1995: 157-159) hold that links between senses can be direct or indirect, depending on whether they are connected by a semantic bridge (i.e. an explicit or implicit semantic component shared between definitions). In the case of metaphorical extensions, a component is introduced in the definition of a derived sense to serve as a semantic bridge with the source sense (ibid: 161).

Arguing against treating sense links merely as an appendage to definitions, Barque (2008: 84-85) dedicated a separate module to represent sense links in a dictionary project. Thanks to this new approach, the researcher refined Martin (1979)'s typology of sense links and characterized them more systematically. The inclusion relationship between senses is analyzed in terms of the central

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1949 (Article)

or peripheral status of the shared semantic component in the definitions. The actantial structure and referential nature (abstract or concrete) of polysemic senses are compared to determine whether and how they change due to meaning extension. Additionally, the rhetorical effect (contiguity or analogy) or the lack thereof is also examined in modeling sense links (Barque 2008: 117). By considering the syntactic, semantic, and rhetorical dimensions, the researcher identified four types of sense links: restriction, extension, metonymy and metaphor. Each category is further divided: restriction is sorted into specialization and euphemism, and extension into generalization and exaggeration. Metonymy is characterized as strong or weak, and metaphor as sensory or structural (ibid: 127).

3.2 CL-informed representation of sense links

According to Geeraerts (2001: 7), cognitive semantics has added a number of new insights to the description of sense links. Prototype Theory characterizes polysemy as follows: one sense may directly or indirectly form the basis of others, carrying more structural weight and functioning as the prototype. Peripheral meanings are derived from, and clustered around the prototypical meaning. All meanings of polysemous lexical items are structured into radial sets and interrelated through family resemblance (Lewandowska-Tomaszczyk 2007: 148). Cognitive semantics has identified various types of motivational links between polysemic senses: metaphor, metonymy, specialization, generalization, profile shift, and image-schema transformations (Radden and Kövecses 1999, Taylor 2003, Tyler and Evans 2004, Gries 2015: 474). These links, grounded in experience, cannot be adequately explained without drawing on language users' experience of their physical, social and cultural surroundings (Boers and Lindstromberg 2008, 2009).

The cognitive linguistic view of polysemy has sparked interest in the pedagogical value of logical sense ordering in learner's dictionaries (Van der Meer 2004, Wojciechowska 2012, Ostermann 2015: 321, Xu and Lou 2015: 224, etc.). Logical sense ordering arranges senses at two levels (i.e. a central or basic level for core senses and a subordinate level for sub-senses) (Moerdijk 2003: 286) and nests subsenses under their corresponding core senses. Due to its linear layout, logical sense ordering cannot fully capture the multidimensional structure of polysemy. To solve the linearization issue identified by Geeraerts (1990: 198), Lu and Wei (2019) proposed a graphic representation of the polysemic structure as a supplement to the linear layout of senses. Instead of a radial network, Lu et al. (2020) presented a left-to-right mind map, where all senses are reduced to short definitions and expressed as nodes, with the prototypical sense placed at the leftmost part and extending rightward to peripheral senses.

Another weakness of logical sense ordering is its failure to explain the extension of one sense to another. To address this problem, some researchers suggested using core definitions to "cover in a general way all derived subsense definitions" (Van der Meer 2000; Smirnova 2016). Halas (2016: 136) proposed

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1949 (Article)

incorporating the dominant semantic component shared with the superordinate sense in the definition of a sub-sense. While these defining strategies sometimes successfully clarify links between core senses and sub-senses, they encounter difficulties when metaphor is involved. An alternative solution is to label senses as "metaphoric extensions" (Smirnova 2016) or use phrases like "resemble" (Halas 2016: 137), "as if" (Van der Meer 2000: 426), and "metaphorized into" (Zhao 2003: 186) to introduce metaphor in definitions. However, this kind of dictionarese may not be clear enough for users to understand the mechanism. Full-sentence definitions (Hanks 1987: 119; Rundell 2006: 324; Atkins and Rundell 2008: 441) offer a remedy to explain sense links (Lu and Wei 2019; Lu et al. 2020). They consist of two parts: the left-hand part introduces the headword, and the right-hand part relates two vertically adjacent senses in the hierarchy. If a sense link is metonymic, the short definition of the superordinate sense is treated as an adverbial or modifier and attached to the subordinate sense. In the case of metaphoric links, a *like* phrase or an *as-if* clause is used to relate the two senses. Occasionally, life experience is invoked to clarify an obscure relation involving conceptual metaphor.

4. Domain specificity and polysemy in specialized dictionaries for learners

Polysemy is notably less common in SDLs than in GDs. For instance, Bergenholtz and Kaufmann (1997) found that out of 2,500 dictionary articles in a dictionary of biotechnology, only three have more than one meaning. Even when the same word is entered in both types of dictionaries, the ratio of meaning to lexical items is still lower in SDLs than in GDs (see Cooper 2005; L'Homme 2020a). This difference can be partly explained by SDLs' usual preference for polysemy that is specific to a particular subject field only and often limited in number compared with other types of polysemy. While it is reasonable for SDLs to focus on a single subject field, it is also possible to reconcile domain specificity with polysemy that spans multiple subject fields or includes both specialized and general meanings.

4.1 Reasons for SDLs' focus on a single field

According to Bergenholtz and Tarp (1995: 59), a specialized dictionary can cover either an entire subject field, several subject fields, or one or more sub-fields, referred to as single-field, multi-field and sub-field dictionaries, respectively. There has been some debate among lexicographers regarding the disciplinary coverage of SDLs. Zhang (2009: 32), for example, believes that SDLs at the initial stages should not be too narrow in their coverage of the subject field. Instead, comprehensive or multidisciplinary dictionaries should be compiled first, and then gradually move towards single-field dictionaries. In contrast, Gouws (2010: 66) argues against covering multiple fields in one SDL, stating it

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1949 (Article)

"may be confusing to the users, especially if each central list text has its own front and back matter texts, constituting a range of secondary frame structures". We also argue against covering multiple fields in SDLs.

Similar to Gouws (2010: 66), our first reason is also concerned with the lexicographer's perspective. Bergenholtz and Tarp (1995: 59) pointed out that multi-field dictionaries are not recommended. They detailed the difficulties that the coverage of multiple subject fields might cause in the compilation process: it is hard to ensure a uniform treatment of the subject fields. For example, lemma selection for a multi-field dictionary is often based on the most frequently used terms or the basic vocabulary of the subject fields. Despite having specialized corpora for some disciplines, lexicographers still need to consult a wide range of experts for lemma selection. Unfortunately, "experts may turn out to differ widely as to what should be considered important or central in their respective subject fields", resulting in "different criteria being employed for practical lemma selection in the same dictionary" (ibid: 60). Another problem with dealing with multiple subject fields simultaneously is related to the treatment of encyclopedic information in SDs. Firstly, the coverage of vocabulary in multi-field dictionaries is often so massive that there is little or no space left for encyclopedic information, which is often necessary for disambiguating terms that may have different meanings across subject fields. Secondly, the preparation of encyclopedic notes requires the involvement of experts from various subject areas, leading to coordination challenges similar to lemma selection. Finally, it is difficult to offer an encyclopedic section or subject-field introduction that provides an overall view of the individual subject areas. Although considered important for the pedagogical dimension of specialized dictionaries (Tarp 2005) and beneficial to layman users in particular (Bergenholtz and Nielsen 2006: 290), a subject-field component covering all subject fields would be too voluminous and complex to be implemented in compilation.

Our second piece of evidence comes from Terminology and relates to the facilitation of learners' acquisition of specialized knowledge of concepts. Terminologists have found that many concepts can be classified in more than one way because there are multiple characteristics that can be used to distinguish between the concepts. This was referred to as multidimensionality by Bowker (1993) in Terminology. This term has since been expanded to mean the "phenomenon where the same concept can be conceptualized from different perspectives" (L'Homme 2020b: 89). Subject fields provide important perspectives that can influence how concepts are related to other concepts. For example, in Engineering, the most prominent conceptual relations to the concept "water" are MADE_OF and AFFECTS, whereas in Geology, CAUSES and TYPE_OF are the most salient conceptual relations. Additionally, subject fields also shape the conceptual categories that a concept can be associated with. In Engineering, "water" is only linked to artificial entities or processes (PUMPING, CONCRETE, CULVERT), while in GEOLOGY it is primarily associated with natural ones (EROSION, GROUNDWATER, SEEPAGE) (León Arauz and Faber 2010). While multidi-

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1949 (Article)

mensionality can be used to enrich traditional static representations in terminological resources, it can also lead to information overload, which hinders knowledge acquisition. This can also be illustrated with the concept "water". According to León Arauz and Faber (2010), 'water' is a versatile concept involved in numerous environment-related situations. Therefore, a large number of conceptual relations will form around "water" if all its dimensions are reflected in the conceptual network. Obviously, users would not acquire meaningful knowledge if they are overwhelmed by a multitude of conceptual relations. The problem of information overload can be solved through recontextualization, such as by specifying a certain subject field. For example, when the contextual constraint of Engineering is applied (León-Araúz et al. 2013: 46), relevant relations (e.g. WATER part_of CONCRETE) will be retained while irrelevant ones (e.g. WATER affects SEEPAGE, which is more typical of Geology) will be filtered out. Recontextualization not only reduces interference from other subject fields but also increases coherence within the specified domain, thereby enhancing the effectiveness of knowledge acquisition.

4.2 Guidelines for treating polysemy in SDLs

The subject coverage of specialized dictionaries is primarily reflected in the selection of headwords. To maintain a focus on a particular subject field, only lexical items specific to the field will be considered for inclusion in the lemma list of an SDL. Since it is usually the meanings, rather than the forms, of lexical items that indicate their affinity with a subject field, the domain specificity of meanings is often used as the criterion for determining whether a word or phrase should be included. However, if the criterion of domain specificity is strictly applied, SDLs will only record polysemes consisting of meanings that are specific to a subject field, as is the case with most specialized dictionaries. This would limit SDLs' ability to utilize sense links to assist learners in acquiring specialized senses they may struggle with. To address this, we propose the following guidelines for handling polysemy in SDLs while still maintaining a focus on a subject field. We will demonstrate these guidelines using computing-related expressions.

4.2.1 The polyseme considered for inclusion in SDLs should contain at least one domain-specific meaning. This guideline establishes the minimum requirement that polysemy must meet in order to be considered in SDLs. As mentioned earlier, the domain specificity of a lemma is represented by the meaning which belongs to a specific subject field. Therefore, a polysemous expression can be included in the lemma list as long as it carries a meaning that is specific to the subject field that defines the disciplinary boundary of an SDL. The requirement for domain specificity should not obscure the fact that the meanings of a polysemous word are often not limited to a single subject field. The composition of polysemic meanings in SDLs can be summarized in the following three

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1949 (Article)

situations: (1) Polysemous terms are exclusively domain-specific, comprising meanings that are unique to a particular subject field (hereafter referred to as domain-specific meanings). For instance, "write-protect" carries two domain-specific meanings: "protect (a disk) from accidental writing or erasure" and "able to stop data being written to or erased from a disk". (2) Polysemes have specialized meanings only, with some being domain-specific and others pertaining to diverse subject fields. For example, the two meanings of "working memory" are specific to the subjects of psychology and computing respectively. (3) Certain polysemic meanings are specific to a particular domain, while others are used in language for general purposes. A typical example is "menu", for which the computing meaning is a metaphorical extension of its general language use. Intra-domain polysemy and inter-domain polysemy, as defined by Meyer and Mackintosh (2000), will be used to designate the first and second types of polysemy respectively. The third type, where the meaning range of polysemy extends beyond specialized domains, will be referred to as extra-domain polysemy.

4.2.2 At least one domain-specific meaning should be addressed as the learning target. This guideline emphasizes the pedagogical considerations when incorporating polysemy in SDLs. When we classify polysemy into three categories, we take a synchronic perspective. However, the three types of polysemy can be understood as resulting from a diachronic process involving semantic extensions between pairs of meanings. The composition of polysemy in SDLs shows that there are three types of meanings: domain-specific meanings, meanings related to other subject fields, and meanings used in language for general purposes. Theoretically, each type of meaning can derive from, and extend to, other types of meanings. Since our goal is to enhance learners' acquisition of domain-specific senses, we are primarily interested in links leading to a domain-specific sense. The domain-specific sense that is to be learned will hereafter be referred to as the target sense, as opposed to the source sense from which it derives. This does not rule out the possibility that a domain-specific sense functioning as the source sense may extend to another domain-specific sense and facilitate its learning. Therefore, the three processes of polysemization described in Section 2.2 will be treated differently: terminologization (as in "menu") will be fully considered in the treatment of polysemy in SDLs but determinologization (such as the computing meaning of "real time" being extended to describe processes like reporting and decision-making) will not be included in meaning descriptions. Re-terminologization will only be taken into account when it results in domain-specific meanings. For instance, the link between the two senses of "working memory" will be considered in SDLs because the psychological sense extends to the computing sense, not the other way around.

4.2.3 The target sense should be explained in relation to the source sense(s) in the meaning description. This guideline clarifies the position of a target sense

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1949 (Article)

relative to its source senses in SDL's representation of polysemy. Semantic extension involves at least two senses: a source sense and a target sense. It is important to have an operational description of both terms. According to L'Homme and Polguère (2008), the perception of semantic extension is usually connected to a diachronic reality. They argue that "it is necessary to decide whether one bases oneself on the true etymology, as one can retrace it in a historical dictionary, or on the intuitive perception of the ordinary speaker" (ibid). We will adopt the second approach: two senses will be treated as the source sense and the target sense, respectively, as long as the meaning description of the latter can build on the former, regardless of the chronological sequence of their earliest occurrences as documented in a historical dictionary. For example, citations in *The Oxford English Dictionary* (Simpson and Weiner 1989) show that the intransitive use of "reboot" is younger (more recent) than the transitive use. However, the former is still considered the source sense and the latter the target sense because the transitive use specifies the agent by building on the argument structure of the intransitive use and the arrangement aligns with ordinary people's intuition that a syntactically simpler meaning appears first and develops into a more complex one later. The target and source senses are not of equal importance in the meaning description: the latter is a means to an end, i.e. it is used to help learners acquire the former. Therefore, they are treated differently at the microstructural level: the target sense always receives a definition in SDLs. In contrast, the source sense is usually not defined unless it is specific to the same domain. In terms of the meaning description of a given target sense, the source sense is either used as part of the gloss for the definition of the target sense (e.g., placed within brackets to make explicit the sense link) or incorporated into its definition (e.g., embedding the meaning of an inchoative verb in that of a causative verb).

5. Representation of sense links in bilingual specialized dictionaries for ESP learners

Building sense links involves bridging semantic gaps between polysemic senses. Therefore, the magnitude of the semantic gap, or the semantic distance, must be assessed to determine how sense links are represented in BSDLs. There have been attempts to characterize the semantic distance between polysemic senses. For instance, Mel'čuk and Polguère (1995: 162-171) categorize it as large, medium and small, depending on the extent of the semantic intersection and the regularity of the semantic distinction. L'Homme (2020b: 107) suggests that the scale of semantic distance accounts for three forms of polysemy: long-distance polysemy, which occurs between a basic meaning and a metaphorical extension; short-distance polysemy, where one or a few semantic components are shared by the lexical units; and regular polysemy, as originally defined by Apresjan (1974: 16). These discussions provide useful insights, but they only address monolingual settings. When representing sense links in BSDLs, the linguistic dimension must also be considered because the perceived closeness between polysemic senses in a bilingual dictionary may be influenced by the degree of overlap between their target language (TL) equivalents.

In what follows, we will illustrate the lexicographical representation of sense links using computing-related polysemy extracted from *The English–Chinese Dictionary (Unabridged)* (Lu 2007). Thanks to the subject labels for computer science, we were able to retrieve all polysemous items with at least one computing-specific sense. When phrasing definitions, we also drew upon specialized dictionaries such as the *Oxford Dictionary of Computing for Leaners of English* (Pyne and Tuck 1996) and the *Dictionnaire fondamental de l'informatique et de l'internet* (L'Homme 2024). By surveying and adapting the extracted data, we will attempt to represent sense links in BSDLs. The proposed model takes into account factors such as the semantic intersection of the polysemic senses, the regularity of their semantic distinction, the mechanism for semantic extension, and the overlap between their Chinese equivalents. Since we are only interested in information categories that facilitate the explanation of sense links, we will leave out pronunciations but retain parts of speech, subject labels, and TL equivalents. Illustrative sentences will be omitted to highlight the layout of the model, although they are particularly useful in describing the meaning of predicative senses. Moreover, to make the overlap between the source and target senses identifiable and accessible, we will italicize the shared parts of their Chinese equivalents. According to the lexicographical devices needed to bring out the connections between senses, the models are presented using one of the following four means: ordering senses logically, appending the source sense, combining logical ordering with short explanation, or providing the source sense and a short explanation. We will also translate the right-core semantic comment of each entry into English in a literal (and perhaps unnatural) way to help non-Chinese readers of this paper understand how the Chinese equivalents overlap and how two senses in Chinese are linked (Please note that the English translation is not intended as part of the BSDL models).

5.1 Lexicographical representation by ordering senses logically

When the senses forming regular polysemy are all domain-specific and their TL equivalents clearly overlap, links between them can be represented by placing the source sense before the target sense.

(1) GIF n.【计】<computing>
  1. *图形交换格式* <*graphic interchange format*>
  2. *图形交换格式*文件 <a file in *graphic interchange format*>

(1) is a typical case of regular polysemy since the "format to file" pattern of extension can be observed in other polysemous words such as PDF and JPEG. As shown by the Chinese equivalents, the source sense "*图形交换格式*" is in-

cluded in the target sense "*图形交换格式*文件". The former denotes a file format whereas the latter refers to the file in this format. Learners should be able to understand the link between the two senses by comparing their Chinese equivalents. Therefore, there is no need for further lexicographical devices.

(2) boot up【计】<computing>
1. vi.（*电脑、系统*）*启动* <(*computer, system*) *start*>
2. vt.（用户）*启动*（*电脑、系统*）<(user) *start* (*computer, system*)>

(2) illustrates another type of regular polysemy, which is also called inchoative/causative alternation (L'Homme 2020b: 108). In the source sense, the phrasal verb "boot up" is used intransitively, meaning that the computer or system starts by itself. In the target sense, the same expression is used transitively, where the user is the subject who causes the computer or system to start. Although they differ in the argument structure, they are clearly derived from the same underlying event of rebooting, with one realizing part of the event structure linguistically and the other encoding the whole. Therefore, placing the intransitive sense before the transitive one should suffice to account for the sense link.

(3) telnet【计】<computing>
1. n. *远程登录服务* <*remote log-in service*>
2. vi. (访客、用户) 使用*远程登录服务* <(visitor, user) use *remote log-in service*>

(3) results from a process of word-formation traditionally known as conversion. This mode of word-formation is now recategorized by some cognitive semanticists as a process whereby a salient participant is singled out as the "metonymic focus" to designate the whole event (Dirven 1999: 280, Dirven and Verspoor 2004: 64). The new perspective is reflected by the close link between the Chinese equivalents of the two senses, where "telnet" is used to designate the event of using "telnet".

5.2 Lexicographical representation by appending the source sense

When the source sense is not domain-specific and the TL equivalents of the headword in the source and target senses clearly overlap, links between them can be represented by appending the source sense. The added sense will be marked with an arrow pointing to the target sense and placed in the brackets following the target sense.

(4) mouse n.【计】<computing> *鼠*标 [←*鼠*；*老鼠*] <*mouse* pointer [←rat, *mouse*]>

In (4), the source sense "*鼠; 老鼠*" is added to help learners understand why a word often referring to a small rodent can be used in computing to mean an input device especially for a computer. The source sense is terminologized into the target sense through a metaphor that is based on similarity in shape: a classic computer mouse has a cable extending from one end of its grip portion, resembling a mouse dragging its tail behind it. Even if the shape of a mouse has evolved in response to new technology (e.g. a wireless mouse does not have a cable), the shape of the grip portion remains largely the same. Therefore, italicizing the shared character "鼠" contributes to learners' understanding of the sense link.

(5) validation n.【计】<computing> (计算机用户对数据、文件的) *确认* [←批准；*确认*]
<(user's) *confirmation* (of data, file) [←approval, *confirmation*]>

(5) is another example of extra-domain polysemy. The source sense is translated into two Chinese words (i.e. 批准 and 确认), which are combined to cover the meaning. Of the two words, 确认 is used as the Chinese equivalent of the technical sense after it is modified by a phrase specifying the possible agent and patient of the action denoted by the headword "validation". A comparison of the equivalents reveals that the technical sense is actually a specialization of the general sense or a microsense (See Cruse 1995).

5.3 Lexicographical representation by logical ordering plus short explanation

When the source sense and target sense(s) are both domain-specific and there is little or no overlap between their TL equivalents, it is not sufficient to place the source sense before the target sense only. A short explanation placed in the brackets following the target sense is also needed to make explicit its link with the source sense.

(6) bit n.【计】<computing>
1. *二进制位*，二进制数字 <*binary digit*, binary number>
2. 比特（度量信息的最小单位）[比特（bit的音译）用于度量以*二进制位*编码的信息量] <pi-tê (the smallest unit of information) [pi-tê (the transliteration of bit) is used to measure the amount of information encoded in *binary digits*]>

(6) is a case of regular polysemy where one sense denotes a concrete element while the other refers to an abstract measure. However, this element-to-measure link between the two senses of the word "bit" is not obvious to learners because there is no overlap between the equivalents "二进制位，二进制数字" and "比特". As shown by the Chinese translations, the TL equivalents of the source sense are meaning-based whereas that of the target sense is form-based (i.e. transliteration). It is, therefore, necessary to add a short explanation to make explicit the link between the two senses.

(7) initialize vt. 【计】<computing>
1. *预置* <*prepare*>
2. 格式化 (磁盘) [格式化磁盘就是*预置*磁盘，以存储和读取数据] <format (computer disk) [to format a computer disk is *preparing* it for storing and reading data]>

(7) falls into the category of intra-domain polysemy. Of the two computing-specific senses, the first can be defined as "to prepare a piece of computer equipment or software for use" and the second as "to prepare a computer disk for use so that it can store and read data". The semantic intersection between the English definitions shows that the target sense is a specialized case of the source sense. However, the semantic connection is "lost" in translation due to the lack of overlap between their Chinese equivalents. A short gloss, therefore, is provided to restore the sense link.

5.4 Lexicographical representation by providing the source sense and a short explanation

When the source sense is not domain-specific and the overlap between the TL equivalents of the headword in the source and target senses is not sufficient to explain the links between them, a short explanation in addition to the source sense should be provided to make explicit its link with the target sense.

(8) toolbox n.【计】<computing> *工具箱* [←（由木头、塑料或金属制成的）*工具箱*: 可从一个选单调用的一组程序或功能，如同装进*工具箱*的一套工具] <*tool case* [←(wooden, plastic or metal) *tool case*: the set of programs or functions accessible from a single menu is like a set of tools kept in a *tool case*]>

(8) is a case of extra-domain polysemy. The technical sense "the set of programs or functions accessible from a single menu" is a metaphoric extension of the non-specialized sense "a container for keeping tools in". Both senses are translated into 工具箱, resulting in complete overlap between the Chinese equivalents and indicating a link between the two senses. Because it is not easy for learners to connect a sense about "a feature of a program" to one about "a container for tools", a short explanation is provided to highlight the similarity between a menu "containing" a set of programs or functions and a case containing a set of tools.

(9) Winchester n. 【计】<computing> *温切斯特*磁盘 [←*温切斯特*连发步枪: *温切斯特*磁盘按原设计可容纳2个30兆字节的磁盘，其IBM编号为3030，恰与*温切斯特*连发步枪用0.30格林火药的0.30口径子弹相同] <*Winchester* disk [←*Winchester* rifle: The *Winchester* disk, as originally designed, can hold two 30-megabyte disks. Its IBM designation is 3030, coincidentally matching the caliber of the 0.30 cartridge used in the *Winchester* rifle, which fires 0.30 caliber bullets.]>

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1949 (Article)

Similar to (8), the overlap (i.e. the shared name 温切斯特) between the Chinese equivalents of the computing sense and the added sense indicates, rather than explicates, the link between them. It is, in fact, the shared number 3030 that connects the two senses though its meaning in one sense is different from that in the other. However, this etymological knowledge is probably beyond lay people as well as some professionals. For this reason, a short gloss is used to provide learners with the fun fact about the coincidence between the two senses.

(10) syntax n.【计】<computing> *句法* [←【语】*句法*；语法；句子结构(分析): 编程用的指令系统比作是语言，编程的规则因而比作是*句法*, 参见 PARSE, TRANSLATE, DICTIONARY 等词] <*syntax* [← <linguistics> *syntax*; grammar; (analysis of) sentential structure: the instruction systems used for programming are likened to language and, accordingly, the rules of programming are compared to *syntax*. See PARSE, TRANSLATE, DICTIONARY, etc.]>

(10) is an instance of inter-domain polysemy, where the computing sense is reterminologization — or rather, a metaphorical extension of the added linguistic sense. As in (5), the target sense is translated using one of the Chinese equivalents of the source sense. Despite the shared equivalent 句法, learners might find it still difficult to make sense of the similarity between the two technical senses. To enable learners to benefit further from the gloss, we invoke two mappings (i.e., from the language to programming instructions and from syntax to programming rules) of the conceptual metaphor PROGRAMMING IS USING THE WRITTEN FORM OF A HUMAN LANGUAGE without resorting to linguistic jargon. Related terms are cross-referenced to reinforce the impression about the semantic regularity in these lexical items.

6. Concluding remarks

Sense links used to receive little attention in SDLs but can now be further discussed thanks to some favorable changes. For example, Descriptive Terminology has identified more polysemy within a domain than Prescriptive Terminology and removed the restriction of domain specificity to expand polysemy beyond a given domain. Interesting attempts have also been made to represent sense links systematically in formalized lexicons or demonstrate their pedagogical value in learners' dictionaries. Drawing upon insights from these studies, we proposed three guidelines and four models. Specifically, when treating intra-domain polysemy in BDSLs, lexicographers should place the source sense before the target sense and sometimes append a gloss to the target sense to explicate an obscure sense link. When dealing with inter- or extra-domain polysemy, the BDSL's focus on a single subject field must be maintained. The non-domain specific source sense needs to be added as a gloss and used as background knowledge to facilitate the understanding of the target, domain-specific sense. When it is difficult to relate the two senses, a short explanation is included in the gloss to make explicit the link between them.

Our research is useful in the following three ways. First, it demonstrates the feasibility of including polysemy extensively in SDLs that are supposed to be single-field dictionaries. Due to lexicographers' concern about domain specificity, there have been few attempts to exploit sense links in SDLs than in general dictionaries for learners. The present study offers practical suggestions on how to choose and treat polysemy without losing the SDL's focus on a single subject field. Second, our research incorporates the overlap between the TL equivalents of the source and target senses into the description of sense links. Sense links have been characterized chiefly in terms of the semantic intersection of the source and target senses, the regularity of their semantic distinction, and the mechanism for semantic extension. The present research proposes that the overlap between equivalents affects the perceived semantic distance between source and target senses and should be fully considered in describing sense links. Third, the study illustrates how sense links can be treated in BSDLs using computing-related polysemy. Sense links used to be represented by logical sense ordering alone. They are now further described in some research (e.g. Lu and Wei 2019; Lu et al. 2020) by means of definitions carefully crafted to reveal the shared semantic components between the source and target senses. However, these strategies are designed for non-specialized polysemy in monolingual dictionaries. They are therefore adapted to bilingual dictionaries, varied in line with the types of polysemy and embodied in four models.

Nevertheless, there is still much work to be done in advancing the study. Building on previous research in terminology and lexicography, we have proposed models for representing sense links in BSDLs with the aim of aiding learners in acquiring technical senses. Consequently, this approach primarily involves speculation. Therefore, it is necessary to conduct empirical investigations to assess the effectiveness of the models and gather feedback from users to enhance them. Additionally, the success of these models relies on the assumption that dictionary users already possess knowledge of the source sense and can utilize it as a foundation for learning the target sense. While this may hold true for extra-domain polysemy, it is less probable in the case of intra- or inter-domain polysemy, as the source sense itself is technical in nature. Hence, knowledge about morphology, etymology and even mnemonics could prove highly beneficial in helping users to grasp the source sense initially. Lexicographical research in these areas is scarce but extremely valuable for improving BSDLs.

Acknowledgements

This research was supported by the National Social Science Fund of China (No. 21FYYB004). We would like to express our gratitude to the anonymous reviewers for their insightful comments, which have greatly contributed to enhancing the quality of this article.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1949 (Article)

Endnote

1. Source sense is used in this paper to designate a sense that functions as the basis and extends to a specialized sense (which is called target sense). A source sense can be the basic or core sense of a lexical item or an extension of the basic sense.

References

A. Dictionaries

L'Homme, M.-C. (Ed.). 2024. *Dictionnaire fondamental de l'informatique et de l'internet (DiCoInfo).* https://olst.ling.umontreal.ca/dicoinfo/dicoinfo-bilingue-en.html

Lu, G. (Ed.). 2007. *The English–Chinese Dictionary (Unabridged).* Shanghai: Shanghai Translation Publishing House.

Pyne, S. and A. Tuck (Eds.). 1996*. Oxford Dictionary of Computing for Leaners of English.* Oxford: Oxford University Press.

Simpson, J.A. and E.S.C Weiner (Eds.). 1989. *The Oxford English Dictionary.* Second Edition. Oxford: Clarendon Press.

B. Other literature

Apresjan, J. 1974. Regular Polysemy. *Linguistics* 142: 5-32.

Atkins, B.T.S. and M. Rundell. 2008. *The Oxford Guide to Practical Lexicography*. Oxford/New York: Oxford University Press.

Barque, L. 2008. *Description et formalisation de la polysémie régulière du français.* Unpublished Doctoral Dissertation. Paris: Université Paris 7.

Beréndi, M., S. Csábi and Z. Kövecses. 2008. Using Conceptual Metaphors and Metonymies in Vocabulary Teaching. Boers, F. and S. Lindstromberg (Eds.). 2008. *Cognitive Linguistic Approaches to Teaching Vocabulary and Phraseology*: 65-100. Berlin: De Gruyter Mouton.

Bergenholtz, H. and U. Kaufmann. 1997. Terminography and Lexicography: A Critical Survey of Dictionaries from a Single Specialised Field. *Hermes* 18: 91-125.

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 (Eds.). 1995. *Manual of Specialised Lexicography: The Preparation of Specialised Dictionaries*. Amsterdam/Philadelphia: John Benjamins.

Blank, A. 2003. Polysemy in the Lexicon and in Discourse. Nerlich, B., Z. Todd, V. Herman and D.D. Clarke (Eds.). 2003. *Polysemy: Flexible Patterns of Meaning in Mind and Language*: 267-296. Berlin/New York: Mouton de Gruyter.

Boers, F. and S. Lindstromberg. 2008. How Cognitive Linguistics Can Foster Effective Vocabulary Teaching. Boers, F. and S. Lindstromberg (Eds.). 2008. *Cognitive Linguistic Approaches to Teaching Vocabulary and Phraseology*: 1-61. Berlin: De Gruyter Mouton.

Boers, F. and S. Lindstromberg. 2009. *Optimizing a Lexical Approach to Instructed Second Language Acquisition*. Basingstoke: Palgrave Macmillan.

Bowker, L. 1993. Multidimensional Classification of Concepts for Terminological Purposes. Smith, Philip J., Clare Beghtol, Raya Fidel and Barbara H. Kwasnik (Eds.). 1993. *Proceedings of the 4th ASIS SIG/CR Classification Research Workshop Held at the 56th ASIS Annual Meeting, October 24–28, 1993, Columbus, Ohio*: 39-56. Columbus, Ohio: American Society for Information Science.

Cabré, M.T. 1999. *Terminology: Theory, Methods, and Applications.* Amsterdam/Philadelphia: John Benjamins.

Chung, T.M. and P. Nation. 2003. Technical Vocabulary in Specialised Texts. *Reading in a Foreign Language* 15(2):103-116.

Cooper, M. 2005. A Mathematical Model of Historical Semantics and the Grouping of Word Meanings into Concepts. *Computational Linguistics* 32(2): 227-248.

Costa, R. 2013. Terminology and Specialised Lexicography: Two Complementary Domains. *Lexicographica* 29: 29-42.

Cruse, D.A. 1995. Polysemy and Related Phenomena from a Cognitive Linguistics Viewpoint. Saint-Dizier, P. and E. Viegas (Eds.). 1995. *Computational Lexical Semantics*: 33-49. Cambridge: Cambridge University Press.

Csábi, S. 2004. A Cognitive Linguistic View of Polysemy in English and its Implications for Teaching. Achard, M. and S. Niemeier (Eds.). 2004. *Cognitive Linguistics, Second Language Acquisition, and Foreign Language Teaching*: 233-256. Berlin/New York: Mouton de Gruyter.

Darmesteter, A. 1886. *The Life of Words as the Symbols of Ideas.* London: Kegan Paul, Trench & Co.

Dirven, R. 1999. Conversion as a Conceptual Metonymy of Event Schemata. Panther, K.-U. and R. Günter (Eds.). 1999. *Metonymy in Language and Thought*: 275-288. Amsterdam/Philadelphia: John Benjamins.

Dirven, R. and M. Verspoor (Eds.). 2004. *Cognitive Exploration of Language and Linguistics.* Cognitive Linguistics in Practice. Vol. 1. Amsterdam/Philadelphia: John Benjamins.

Felber, H. 1984. *Terminology Manual*. Vienna: Infoterm.

Geeraerts, D. 1990. The Lexicographical Treatment of Prototypical Polysemy. Tsohatzidis, S.L. (Ed.). 1990. *Meanings and Prototypes: Studies in Linguistic Categorization*: 195-210. New York: Routledge.

Geeraerts, D. 2001. The Definitional Practice of Dictionaries and the Cognitive Semantic Conception of Polysemy. *Lexicographica* 17: 6-21.

Gouws, R.H. 2010. The Monolingual Specialised Dictionary for Learners. Fuertes Olivera, P.A. (Ed.). 2010. *Specialised Dictionaries for Learners*: 55-68. Berlin/New York: De Gruyter.

Gries, S.T. 2015. Polysemy. Dąbrowska, E. and D. Divjak (Eds.). 2015. *Handbook of Cognitive Linguistics*. HSK 39: 472-490. Berlin: Walter de Gruyter GmbH & Co KG.

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.

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/Glasgow: Collins ELT.

Lakoff, G. 1987. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind.* Chicago/London: University of Chicago Press.

León-Araúz, P. and P. Faber. 2010. Natural and Contextual Constraints for Domain-Specific Relations. Barbu Mititelu, V., V. Pekar and E. Barbu (Eds.). 2010. *Proceedings of the Workshop, Semantic Relations. Theory and Applications, 18 May 2010, at the International Conference on Language Resources and Evaluation (LREC) 2010, Malta:* 12-17. Malta: ELRA.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1949 (Article)

León-Araúz, P., A. Reimerink and A.G. Aragón. 2013. Dynamism and Context in Specialized Knowledge. *Terminology* 19(1): 31-61.

Lewandowska-Tomaszczyk, B. 2007. Polysemy, Prototypes, and Radial Categories. Geeraerts, D. and H. Cuyckens (Eds.). 2007. *The Oxford Handbook of Cognitive Linguistics*: 139-169. Oxford/New York: Oxford University Press.

L'Homme, M.-C. 1998. Le statut du verbe en langue de spécialité et sa description lexicographique. *Cahiers de lexicologie* 73(2): 61-84.

L'Homme, M.-C. 2005. Sur la notion de «terme». *Meta* 50(4): 1112-1132.

L'Homme, M.-C. 2015. Predicative Lexical Units in Terminology. Gala, N., R. Rapp and G. Bel-Enguix (Eds.). 2015. *Language Production, Cognition, and the Lexicon*: 75-93. Berlin: Springer.

L'Homme, M.-C. 2020a. Revisiting Polysemy in Terminology. Gavriilidou, Z, M. Mitsiaki and A. Fliatouras (Eds.). 2020. *Proceedings of the XIX EURALEX International Congress: Lexicography for Inclusion, 7–9 September 2021, Virtual. Vol. I*: 415-424. Komotini, Greece: Democritus University of Thrace.

L'Homme, M.-C. 2020b. *Lexical Semantics for Terminology: An Introduction.* John Benjamins.

L'Homme, M.-C. 2024. Managing Polysemy in Terminological Resources. *Terminology* 30(2): 216-249.

L'Homme, M.-C. and A. Polguère. 2008. Mettre en bons termes les dictionnaires spécialisés et les dictionnaires de langue générale. Maniez, F. and P. Dury (Eds.). 2008. *Lexicographie et terminologie: histoire de mots. Hommage à Henri Béjoint:* 191-206. Lyon: Presses de l'Université de Lyon. https://www.researchgate.net/publication/237511405_Mettre_en_bons_termes_les_dictionnaires_specialises_et_les_dictionnaires_de_langue_generale

Li, M. and J.R. Kirby. 2015. The Effects of Vocabulary Breadth and Depth on English Reading. *Applied Linguistics* 36(5): 611-634.

Lu, H. and X. Wei. 2019. Structuring Polysemy in English Learners' Dictionaries: A Prototype Theory-Based Model. *International Journal of Lexicography* 32(1): 20-37.

Lu, H., Y. Zhang and X. Hao. 2020. The Contribution of Cognitive Linguistics to the Acquisition of Polysemy: A Dictionary Entry-Based Study with Chinese Learners of English. *International Journal of Lexicography* 33(3): 306-336.

Maby, M. 2016. *An Investigation of L2 English Learners' Knowledge of Polysemous Word Senses.* Unpublished Ph.D. Thesis. Cardiff: Cardiff University.

Martin, R. 1979. La polysémie verbale, esquisse d'une typologie formelle. *Travaux de linguistique et de littérature* 17: 261-276.

Mel'čuk, I.A. and A. Polguère. 1995. *Introduction à la lexicologie explicative et combinatoire.* Louvain-la-Neuve: Duculot.

Meyer, I. and K. Mackintosh. 2000. When Terms Move into Our Everyday Lives: An Overview of De-terminologization. *Terminology* 6(1): 111-138.

Miao, L. 2015. The Semantic Production and Development of Chinese Learners' Polysemous Words: A Corpus-Based Study. *Modern Foreign Languages* 38(2): 217-226.

Mićović, D.N. and L.V. Beko. 2022. Polysemy-Related Problems in ESP Students — A Case Study. *Зборник радова Филозофског факултета у Приштини* 52(3): 123-144.

Moerdijk, F. 2003. The Codification of Semantic Information. Van Sterkenburg, P. (Eds.). 2003. *A Practical Guide to Lexicography*: 273-296. Amsterdam/Philadelphia: John Benjamins.

Morimoto, S. and S. Loewen. 2007. A Comparison of the Effects of Image-Schema-Based Instruction and Translation-Based Instruction on the Acquisition of L2 Polysemous Words. *Language Teaching Research* 11(3): 347-372.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1949 (Article)

Nation, I.S.P. 1990. *Teaching and Learning Vocabulary*. New York: Newbury House.

Nation, I.S.P. 2013. *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Nerlich, B. and D.D. Clarke. 1997. Polysemy: Patterns of Meaning and Patterns in History. *Historiographia linguistica* 24(3): 349-385.

Ostermann, C. 2015. *Cognitive Lexicography: A New Approach to Lexicography Making Use of Cognitive Semantics.* Lexicographica. Series Maior 149. Berlin/Boston: Walter de Gruyter GmbH.

Radden, G. and Z. Kövecses. 1999. Towards a Theory of Metonymy. Panther, K.-U. and G. Radden (Eds.). 1999. *Metonymy in Language and Thought*: 17-59. Amsterdam/Philadelphia: John Benjamins.

Read, J. 2004. Plumbing the Depths: How Should the Construct of Vocabulary Knowledge Be Defined? Bogaards, P. and B. Laufer (Eds.). 2004. *Vocabulary in a Second Language: Selection, Acquisition and Testing*: 209-227. Amsterdam: John Benjamins.

Rey, A. 1979. *La terminologie: noms et notions.* Paris: Presses universitaires de France.

Richards, J.C. 1976. The Role of Vocabulary Teaching. *TESOL Quarterly* 10: 77-89.

Rundell, M. 2006. More than One Way to Skin a Cat: Why Full-Sentence Definitions Have not Been Universally Adopted. Corino, E., C. Marello and C. Onesti (Eds.). 2006. *Proceedings of the 12th EURALEX International Congress, Torino, Italia, 6–9 September 2006*: 323-337. Alessandria: Edizioni Dell'Orso.

Sager, J.C. 1990. *A Practical Course in Terminology Processing*. Amsterdam/Philadelphia: John Benjamins.

Schlegel, A.W. 1832. *Réflexions sur l'étude des langues asiatiques adressées à Sir James Mackintosh, suivies d'une lettre à M. Horace Hayman Wilson.* Paris: Maze.

Smirnova, A.Y. 2016. "Where is the Bank?" or How to "Find" Different Senses of a Word. *Heliyon* 2(6): e00065.

Swales, J. 1990. *Genre Analysis: English for Academic and Research Settings.* Cambridge: Cambridge University Press.

Tarp, S. 2005. The Pedagogical Dimension of the Well-Conceived Specialised Dictionary. *Ibérica* 10: 7-21.

Taylor, J.R. 2003. *Linguistic Categorization.* New York: Oxford University Press.

Temmerman, R. 1997. Questioning the Univocity Ideal. The Difference Between Socio-Cognitive Terminology and Traditional Terminology. *Hermes* 18: 51-90.

Temmerman, Rita. 2000. *Towards New Ways of Terminology Description. The Sociocognitive-Approach.* Amsterdam/Philadelphia: John Benjamins.

Tyler, A. and V. Evans. 2004. Applying Cognitive Linguistics to Pedagogical Grammar: The Case of *Over*. Achard, M. and S. Niemeier (Eds.). 2004. *Cognitive Linguistics, Second Language Acquisition, and Foreign Language Teaching*. Studies on Language Acquisition 18: 257-280. Berlin/New York: Mouton de Gruyter.

Tyler, A., C. Mueller and V. Ho. 2011. Applying Cognitive Linguistics to Learning the Semantics of English *to*, *for* and *at*: An Experimental Investigation. *Vigo International Journal of Applied Linguistics* 8: 181-205.

Van der Meer, G. 2000. Core, Subsense and the *New Oxford Dictionary of English* (NODE): On How Meanings Hang Together, and Not Separately. Heid, U., S. Evert, E. Lehmann and C. Rohrer (Eds.). 2000. *Proceedings of the Ninth EURALEX International Congress, EURALEX 2000, Stuttgart, Germany, 8–12 August, 2000. Vol I*: 419-431. Stuttgart: Institut für maschinelle Sprachverarbeitung, University of Stuttgart.

Van der Meer, G. 2004. On Defining: Polysemy, Core Meanings and "Great Simplicity". Williams, G. and S. Vessier (Eds.). 2004. *Proceedings of the Eleventh EURALEX International Congress* EURA-

LEX 2004, *Lorient, France, 6–10 July, 2004. Vol. 3*: 807-815. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud.

Van der Meer, G. 2010. The Treatment of Figurative Meaning in Specialised Dictionaries for Learners. Fuertes-Olivera, P.A. (Ed.). 2010. *Specialised Dictionaries for Learners*. Lexicographica. Series Maior 136: 131-139. Berlin/New York: Walter de Gruyter.

Verspoor, M.H. and W. Lowie. 2003. Making Sense of Polysemous Words. *Language Learning* 53(3): 547-586.

Wojciechowska, S. 2012. *Conceptual Metonymy and Lexicographic Representation*. Frankfurt am Main: Peter Lang.

Wüster, E. 1991. *Einführung in die allgemeine Terminologielehre und terminologische Lexikographie.* Bonn: Romanistischer Verlag.

Xu, H. and Y. Lou. 2015. Treatment of the Preposition *to* in English Learners' Dictionaries: A Cognitive Approach. *International Journal of Lexicography* 28(2): 207-231.

Zhang, Y. 2009. Some Considerations on Bilingual Teaching and Bilingual Specialized Learner's Dictionaries. *Foreign Language World* 133(4): 30-37.

Zhao, H., T. Yau, K. Li. and N. Wong. 2018. Polysemy and Conceptual Metaphors: A Cognitive Linguistics Approach to Vocabulary Learning. Tyler, A., L. Huang and H. Jan (Eds.). 2018. *What is Applied Cognitive Linguistics? Answers from Current SLA Research*: 257-286. Berlin/Boston: Mouton de Gruyter.

Zhao, Y. 2003. *Cognitive Exploration of Lexicography*. Shanghai: Shanghai Foreign Language Education Press.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1952

Publikasieaankondigings / Publication Announcements

Annette Klosa-Kückelhaus (Editor). *Internet Lexicography: An Introduction.* 2024, xi + 312 pages. ISBN 978-3-11-120521-2 (Hardback), ISBN 978-3-11-123375-8 (ePDF), ISBN 978-3-11-123392-5 (ePUB). Lexicographica. Series Maior 164. Berlin/Boston: De Gruyter. Open Access Book.

Annette Klosa-Kückelhaus and Martina Nied Curcio (Editors). *Dictionary Use and Dictionary Teaching: New Challenges in a Multilingual, Digital and Global World.* 2024, viii + 301 pages. ISBN 978-311-137-229-7 (Hardback), ISBN 978-311-137-329-4 (eBook). Lexicographica. Series Maior 166. Berlin/Boston: De Gruyter. Open Access Book.

Phillip A. Louw (Hoofredakteur). *Woordeboek van die Afrikaanse Taal, Agtiende Deel: U.* 2024. www.wat.co.za. Stellenbosch: Buro van die WAT. R200 vir 12 maande (Aanlyn WAT).

Hou Min. *A Dictionary of Chinese Neologisms (2000–2020).* 2023, XL + 786 pp. ISBN 978-7-100-21777-4 (Hardback). Beijing: The Commercial Press. Price $17.69. (Review in this issue.)

Antoni Nomdedeu-Rull and Sven Tarp. *Introducción a la lexicografía en español: Funciones y aplicaciones.* 2024, 256 pages. ISBN 9780367523480 (Hardback), ISBN 9780367523503 (Paperback), ISBN 97810030057567 (eBook). London: Routledge. Price: £135.00 (Hardback), £36.99 (Paperback), £36.99 (eBook).

Alfonso Rascón Caballero. *The Theory and Practice of Examples in Bilingual Dictionaries.* 2024, xii + 317 pages. ISBN 978-3-11-137561-8 (Hardback), ISBN 978-3-11-137564-9 (ePDF), ISBN 978-3-11-137564-9 (ePUB). Lexicographica. Series Maior 165. Berlin/Boston: De Gruyter. Price €109.95 (Hardback, ePDF and ePUB).

Heming Yong, Jing Peng and Xiangming Zhang. *Chinese Lexicography in the Twentieth Century.* 2024, x + 280 pages. ISBN 978-1-6366-7529-9 (Hardback), ISBN 978-1-6366-7527-5 (eBook PDF), ISBN 978-1-6366-7528-2 (ePub). New York /Berlin / Brussels / Chennai / Lausanne / Oxford: Peter Lang. Price: US$ 94.95 (Hardback), US$ 90.20 (eBook). (Review in this issue.)

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)

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 enigeen 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 evaluerend 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, Posbus 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. Dít 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**.

http://lexikos.journals.ac.za; https://doi.org/10.5788/34-1-1951 (Volle uitgawe / Full volume)