http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)

Lexikos 32(2)

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)

Lexikos 32(2)

Huldeblyk aan
Tribute to
D.J. Prinsloo

Redakteur
Editor
Elsabé Taljard

AFRILEX African Association for Lexicography

AFRILEX-REEKS 32A:2022
AFRILEX SERIES 32A:2022

BURO VAN DIE WAT
STELLENBOSCH

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)

Uitgewer Publisher

BURO VAN DIE WAT
Posbus 245
7599 STELLENBOSCH

Eerste uitgawe 2022

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

Lexikos 32(2) AFRILEX-REEKS 32A:2022
Lexikos 32(2) AFRILEX SERIES 32A:2022

Gedruk en gebind deur SUN MeDIA Stellenbosch
Office 5, The Woodmill, Vredenburg Road, Stellenbosch

ISBN 978-1-990998-49-2
ISSN 2224-0039

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

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)

Span van Roterende Redakteurs / Team of Rotating Editors

Adviesraad / Advisory Board

Redaksiekomitee / Editorial Committee

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)

Inhoud / Contents

Artikels / Articles

Voorwoord

Dit is nie moontlik om binne die bestek van 'n voorwoord 'n volledige oorsig te gee oor die bydrae wat professor Danie Prinsloo tydens sy loopbaan tot die leksikografie gemaak het nie. Wat volg, is enkele hoogtepunte van veral joernaalartikels wat die leksikografiepraktyk en -teorie in (veral) Suid-Afrika in nuwe rigtings gestuur het; navorsing waarvan ook die internasionale leksikografiewêreld deeglik kennis geneem het.

Prinsloo begin sy akademiese loopbaan as grammatikus: sy MA-verhandeling is getiteld *Lokatiefvorming in Noord-Sotho* (1979), gevolg deur 'n doktorale proefskrif, *Woordvolgorde en volgordeverandering in Noord-Sotho* (1984). Sy publikasies gedurende die tagtigerjare handel dan ook hoofsaaklik oor hierdie twee temas, maar brei later uit na ander taalkundetemas, waaronder pronominalisasie (Prinsloo 1987). Hy begewe hom ook op die terrein van taalassessering, spesifiek die opstel van meervoudige keuse-items, met Noord-Sotho as fokus (Raubenheimer en Prinsloo 1986; Prinsloo en Raubenheimer 1988 en Raubenheimer en Prinsloo 1989).

Die verskyning van die eerste uitgawe van die *Collins Cobuild English Language Dictionary* (CCELD) (1987) kenmerk die begin van die korpusrevolusie in die internasionale leksikografiewêreld. Die eerste teken dat die revolusie van korpus-gebaseerde en korpus-gedrewe leksikografie ook na Suid-Afrika begin oorspoel, is die verskyning van Prinsloo se artikel *Towards Computer-assisted Word Frequency Studies in Northern Sotho* (Prinsloo 1991). Hierdie artikel verteenwoordig die draaipunt in Prinsloo se akademiese loopbaan — hy word onherroeplik 'n leksikografie-dissipel en staan aan die voorpunt van die korpusrevolusie in die Suid-Afrikaanse leksikografie, spesifiek dié van die Afrikatale. In samewerking met plaaslike en internasionale kollegas begin hy korpusse bou vir al die inheemse tale, met die samestelling van korpus-gebaseerde woordeboeke as oogmerk. In 2000 is hy mede-outeur van drie artikels (De Schryver en Prinsloo 2000a, 2000b en 2000c) wat spesifiek afgestem is op die saamstel van elektroniese korpora vir die Afrikatale en die gebruik van hierdie korpora as basis vir die kompilasie van Afrikataalwoordeboeke. Hierdie artikels word gesien as 'n bloudruk vir korpus-gebaseerde woordeboeke in die Afrikatale.

In 1992 verskyn die eerste artikel van 11 oor lemmatiseringstrategieë van geselekteerde woordkategorieë (Prinsloo 1992). In hierdie artikel en ook dié wat volg, verskaf hy telkens 'n kritiese oorsig oor bestaande strategieë en kom dan met innoverende, teoreties goedgefundeerde oplossings vorendag wat die spesifieke problematiek van die Afrikatale aanspreek. Hoewel sy resultate meestal op Noord-Sotho-/Sepedidata gebaseer is, is dit dikwels ook van toepassing op ander Afrikatale, veral dié met 'n disjunktiewe skryfwyse. 'n Besondere bydrae in dié verband is die navorsing oor die lemmatisering van verwantskapsterminologie in Sepedi en Zulu (Prinsloo 2012a en Prinsloo en Bosch 2012). Komplekse verwantskapsterme is 'n kenmerk van die Afrikakultuur en die lemmatisering en leksikografiese behandeling daarvan behoort besondere aandag in woordeboeke te geniet.

Gedurende die middel-negentigerjare tree e-leksikografie en e-woordeboeke sterk na vore in die internasionale leksikografiewêreld, en ook hier trap Prinsloo diep spore. Reeds in 2001 verskyn sy artikel *The Compilation of Electronic Dictionaries for the African Languages* (Prinsloo 2001). In hierdie artikel dui hy aan hoe sommige van die dringendste lemmatiseringsprobleme wat leksikograwe in die Afrikatale ervaar in elektroniese woordeboeke opgelos kan word. Saam met 'n kollega ontwikkel hy die konsep van 'Simultaneous feedback' en later 'Fuzzy SF' wat daarop neerkom dat enige gebruiker toegang het tot 'n pasgemaakte woordeboek waaruit hulle inligting kan onttrek wat op hulle eie behoeftes afgestem is (De Schryver en Prinsloo 2000d en De Schryver en Prinsloo 2001). Hierdie konsep en die implikasies wat dit vir die metodologie vir die saamstel van woordeboeke het, kan met reg as rewolusionêr beskryf word.

Een van die belangrikste bydraes tot die leksikografiepraktyk deur Prinsloo is ongetwyfeld die ontwerp van 'n multidimensionele leksikografiese liniaal (Prinsloo en De Schryver 2002). Hierdie instrument word gebruik in die beplanning van die makrostruktuur van 'n woordeboek ten einde die oor- en/of onderbehandeling van alfabetiese kategorieë te verhoed. Dit word aanvanklik vir Afrikaans ontwerp, maar in latere publikasies word dit verder verfyn en uitgebrei om ook vir Sepedi voorsiening te maak. Die beskikbaarheid van so 'n multidimensionele meetinstrument is veral vir die Afrikataalleksikografie van onskatbare waarde, aangesien die publikasie van papierwoordeboeke steeds die verstekwaarde vir hierdie tale is en dit is juis vir papierwoordeboeke wat hierdie aspekte van belang is.

Sy grondige kennis van die grammatika van Noord-Sotho kom hom goed te pas wanneer hy hom op die terrein van mensliketaaltegnologie begewe. Hy speel 'n belangrike rol in die ontwikkeling en evaluering van speltoetsers vir die Afrikatale (De Schryver en Prinsloo 2004 en Prinsloo en De Schryver 2004). Die ontwikkeling van annoteringstelle en die annotering van tekskorpora is 'n logiese uitvloeisel van sy aanvanklike werk in korpussamestelling. In samewerking met plaaslike en internasionale kollegas ontwikkel hy 'n annoteringstel vir Sepedi, asook strategieë vir die rekenaarmatige identifisering van naamwoorde en werkwoorde in dié taal. In 2015 lewer Prinsloo saam met sy twee seuns 'n referaat getitled *A Writing Tool for Sepedi* by die Elex-kongres in die Verenigde Koninkryk. Die referaat word baie goed ontvang en trek baie aandag onder die hoofsaaklik internasionale gehoor. Die Sepedi Helper, soos wat hierdie instrument genoem word, is gekonspetualiseer as 'n ondersteuningsinstrument vir woordeboekgebruikers wat addisionele hulp met ingewikkelde grammatikale konstruksies nodig het. Die einddoel van hierdie navorsing is die ontwerp en implementering van 'n volledige elektroniese grammatika vir Sepedi wat uiteindelik deel van 'n interaktiewe e-woordeboek sal uitmaak. Wat die Sepedi Helper van ander, soortgelyke skryfhulpsisteme onderskei, is die feit dat die nie (net) op die leksikale vlak funksioneer nie, maar ook op dié van die morfosintaksis. 'n Gebruikerstudie waaroor daar in 'n opvolgartikel (Prinsloo en Taljard 2019) berig word, bevestig dat die Sepedi Helper wel in sy doel slaag: dit bevestig korrekte response, dit korrigeer foute en dit identifiseer die mees algemene foute wat deur die aanleerders van Sepedi gemaak word. Indien dit in ag geneem word dat Sepedi — soos die ander Afrikatale van Suid-Afrika — relatief arm aan elektroniese hulpbronne is, is die ontwerp en implementering van 'n funksionerende skryfhulpinstrument vir dié taal soveel te meer indrukwekkend.

Saam met vermaarde akademikus en mede-leksikograaf Rufus Gouws, publiseer Prinsloo 'n handboek *Principles and Practice of South African Lexicography* (Gouws en Prinsloo 2005), wat 'n onmisbare verwysingsbron vir enige student in die leksikografie is.

Prinsloo se status as internasionaal gerekende leksikograaf word ook weerspieël in die talle hoofstukke in hoogaangeskrewe internasionale publikasies, waaronder *The Routledge Handbook of Lexicography* (Prinsloo, D.J., J.V. Prinsloo en Daniel Prinsloo 2018), *An International Encyclopedia of Lexicography* (Prinsloo 2013) en *eLexicography* (Prinsloo 2012b), om maar enkeles te noem.

Benewens talle referate op internasionale leksikografiekongresse waaronder Asialex, EMLex, LREC en Australex, word hy ook in 2019 uitgenooi om as gasspreker tydens die 13de internasionale Asialex-kongres te Istanbul in Turkye op te tree, gevolg deur 'n soortgelyke uitnodiging vir die 19de EURALEX kongres in Alexandroupoli, Griekeland.

Soos wat Alberts (hierdie uitgawe) aandui, was Danie Prinsloo instrumenteel in die stigting van Afrilex in 1995, met *Lexikos* as die amptelike tydskrif van die organisasie. Dit is daarom gepas dat 'n spesiale uitgawe van *Lexikos* saamgestel is om erkenning te verleen aan die bydrae wat hy tot die leksikografie gelewer het.

Elkeen van die bydraes in hierdie spesiale uitgawe verteenwoordig 'n aspek van die leksikografie waartoe Danie Prinsloo bygedra het. Hy is egter nie net navorser nie; hy is ook gewaardeerde kollega, dosent en studieleier. Ons hoop dat sy aftrede nie ook sy uittrede uit die wêreld van die leksikografie beteken nie; daarvoor is sy kundigheid veels te kosbaar.

Hierdie spesiale uitgawe sou nie moontlik gewees het sonder die ondersteuning van die Buro van die WAT en sy personeel nie. 'n Besondere woord van dank gaan dus aan Willem Botha, Tanja Harteveld en Hermien van der Westhuizen wat hulle soos gewoonlik, uitstekend van hulle taak gekwyt het.

Bibliografie

De Schryver, G.-M. en D.J. Prinsloo. 2000a. The Compilation of Electronic Corpora, with Special Reference to the African Languages. *Southern African Linguistics and Applied Language Studies* 18(1–4): 89-106.

De Schryver, G.-M. en D.J. Prinsloo. 2000b. Electronic Corpora as a Basis for the Compilation of African-language Dictionaries, Part 1: The Macrostructure. *South African Journal of African Languages* 20(4): 291-309.

De Schryver, G.-M. en D.J. Prinsloo. 2000c. Electronic Corpora as a Basis for the Compilation of African-language Dictionaries, Part 2: The Microstructure. *South African Journal of African Languages* 20(4): 310-330.

De Schryver, G.-M. en D.J. Prinsloo. 2000d. The Concept of 'Simultaneous Feedback': Towards a New Methodology for Compiling Dictionaries. *Lexikos* 10: 1-31.

De Schryver, G.-M. en D.J. Prinsloo. 2001. Fuzzy SF: Towards the Ultimate Customised Dictionary. *Studies in Lexicography* (11)1: 97-111.

De Schryver, G.-M. en D.J. Prinsloo. 2004. Spellcheckers for the South African Languages, Part 1: The Status Quo and Options for Improvement. *South African Journal of African Languages* 24(1): 57-82.

Gouws, R.H. en D.J. Prinsloo. 2005. *Principles and Practice of South African Lexicography.* Stellenbosch: SUN PReSS, AFRICAN SUN MeDIA.

Prinsloo, D.J. 1979. *Lokatiefvorming in Noord-Sotho.* Ongepubliseerde M.A.-verhandeling. Pretoria: Universiteit van Pretoria.

Prinsloo, D.J. 1984. *Woordvolgorde en volgordeverandering in Noord-Sotho.* Ongepubliseerde doktorale proefskrif. Pretoria: UNISA.

Prinsloo, D.J. 1987. Perspektief op pronominalisasie in Noord-Sotho. *South African Journal of African Languages* 7(1): 23-33.

Prinsloo, D.J. 1991. Towards Computer-assisted Word Frequency Studies in Northern Sotho. *South African Journal of African Languages* 11(2): 54-60.

Prinsloo, D.J. 1992. Lemmatization of Reflexives in Northern Sotho. *Lexikos* 2: 178-191.

Prinsloo, D.J. 2001. The Compilation of Electronic Dictionaries for the African Languages. *Lexikos* 11: 139-159.

Prinsloo, D.J. 2012a. Die leksikografiese bewerking van verwantskapsterme in Sepedi. *Lexikos* 22: 272-289.

Prinsloo, D.J. 2012b. Electronic Lexicography for Lesser-resourced Languages: The South African Context. Granger, Sylvaine en Magali Paquot (Reds.). 2012. *Electronic Lexicography*: 119-144. Oxford: Oxford University Press.

Prinsloo, D.J. 2013. The Utilization of Bilingual Corpora for the Creation of Bilingual Dictionaries. Gouws, R.H. et al. 2013. *Dictionaries: An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography:* 1344-1356. HSK 5.4. Berlyn/Boston: De Gruyter Mouton.

Prinsloo, D.J. en S.E. Bosch. 2012. Kinship Terminology in English–Zulu/Northern Sotho Dictionaries — A Challenge for the Bantu Lexicographer. Fjeld, Ruth Vatvedt en Julie Matilde Torjusen (Reds.). 2012. *Proceedings of the 15th Euralex International Congress, 7–11 August 2012, Oslo:* 296-303. Oslo: Departement Linguistiek en Skandinawiese Studies, Universiteit van Oslo.

Prinsloo, D.J. en G.-M. de Schryver. 2002. Designing a Measurement Instrument for the Relative Length of Alphabetical Stretches in Dictionaries, with Special Reference to Afrikaans and English. Braasch, A. en C. Povlsen (Reds.). 2002. *Proceedings of the Tenth Euralex International Congress, EURALEX 2002, Copenhagen, Denmark, August 13–17, 2002*: 483-494. Kopenhagen: Center for Sprogteknologi, Københavns Universitet.

Prinsloo, D.J. en G.-M. de Schryver. 2004. Spellcheckers for the South African Languages, Part 2: The Utilisation of Clusters of Circumfixes. *South African Journal of African Languages* 24(1): 83-94.

Prinsloo, D.J., J.V. Prinsloo en D. Prinsloo. 2018. African Lexicography in the Internet Era. Pedro A. Fuertes Olivera (Red.). 2018. *The Routledge Handbook of Lexicography:* 487-502. Londen: Routledge.

Prinsloo, D.J., D. Prinsloo en J.V. Prinsloo. 2015. A Writing Tool for Sepedi. *E-Lex 2015*, Herstmonceux Castle, United Kingdom, 11–13 August 2015.

Prinsloo, D.J. en R.I. Raubenheimer. 1988. Troubleshooting Multiple-choice Items in Northern Sotho Item Banks for First Language Education. *South African Journal of African Languages* 8(3): 93-98.

Prinsloo, D.J. en E. Taljard. 2019. The Sepedi Helper Writing Assistant: A User Study. *Language Matters* 50(2): 73-99.

Raubenheimer, R.I. en D.J. Prinsloo. 1986. The Writing of Multiple-choice Items for Northern Sotho (First Language). *South African Journal of African Languages* 6 (Supplement): 101-135.

Raubenheimer, R.I. en D.J. Prinsloo. 1989. Item Analysis for Improving Multiple-choice Test Items in North Sotho. *South African Journal of African Languages* 9(2): 70-73.

Elsabé Taljard
Redakteur

Foreword

It is not possible to give a complete overview of the contribution that Professor Danie Prinsloo made to lexicography during his career within the scope of a preface. What follows are some highlights of especially journal articles that have taken lexicography practice and theory in (especially) South Africa in new directions; research that the international lexicography world has also taken serious note of.

Prinsloo started his academic career as a grammarian: his MA dissertation was entitled *Lokatiefvorming in Noord-Sotho* [*Locative Formation in Northern Sotho*] (1979), followed by a doctoral thesis, *Woordvolgorde en volgordeverandering in Noord-Sotho* [*Word Order and Order Change in Northern Sotho*] (1984). His publications during the eighties therefore mainly deal with these two themes, but later expanded to other linguistic themes, including pronominalization (Prinsloo 1987). He also ventures into the field of language assessment, specifically the compilation of multiple-choice items, with Northern Sotho as the focus (Raubenheimer and Prinsloo 1986; Prinsloo and Raubenheimer 1988 and Raubenheimer and Prinsloo 1989).

The publication of the first edition of the *Collins Cobuild English Language Dictionary* (CCELD) (1987) marks the beginning of the corpus revolution in the international lexicography world. The first sign that the revolution of corpus-based and corpus-driven lexicography was also beginning to spill over to South Africa is the appearance of Prinsloo's article *Towards Computer-assisted Frequency Studies in Northern Sotho* (Prinsloo 1991). This article represents the turning point in Prinsloo's academic career — he irrevocably becomes a lexicography disciple and finds himself at the forefront of the corpus revolution in South African lexicography, specifically that of the African languages. In collaboration with local and international colleagues, he starts building corpora for all indigenous languages, with the compilation of corpus-based dictionaries as its goal. In 2000 he co-authors three articles (De Schryver and Prinsloo 2000a, 2000b and 2000c) focusing on the compilation of electronic corpora for African languages and the use of these corpora as a basis for compiling African language dictionaries. These articles are seen as a blueprint for corpus-based dictionaries in the African languages.

In 1992, the first of 11 articles on lemmatization strategies of selected word categories is published (Prinsloo 1992). In this article and also in those that follow, he provides a critical overview of existing strategies and comes up with innovative, theoretically well-founded solutions that address the specific problems of the African languages. Although the results are mostly based on Northern Sotho/Sepedidata, they are often also applicable to other African languages, especially those with a disjunctive writing system. A particular contribution in this regard is his research on the lemmatization of kinship terminology in Sepedi and Zulu (Prinsloo 2012a and Prinsloo and Bosch 2012). Complex kinship terms are a feature of African culture and their lemmatization and lexicographical treatment should receive particular attention in dictionaries.

During the mid-nineties, e-lexicography and e-dictionaries come to the fore in the international lexicography world, and here too, Prinsloo makes his mark. Already in 2001 his article *The Compilation of Electronic Dictionaries for the African Languages* appears (Prinsloo 2001). In this article he illustrates how some of the most pressing lemmatization problems experienced by lexicographers in the African languages can be solved in electronic dictionaries. Together with a colleague he develops the concept of 'Simultaneous feedback' and later 'Fuzzy SF' which means that any user has access to a custom dictionary from which they can extract information tailored to their own needs (De Schryver and Prinsloo 2000d and De Schryver and Prinsloo 2001). This concept and the implications it has for the methodology of dictionary compilation can rightly be described as revolutionary.

One of the most important contributions to lexicographic practice by Prinsloo is undoubtedly the design of a multidimensional lexicographic ruler (Prinsloo and De Schryver 2002). This tool is used in planning the macrostructure of a dictionary in order to prevent the over- and/or undertreatment of alphabetical categories. It was initially designed for Afrikaans, but in later publications it is further refined and expanded to also make provision for Sepedi. The availability of such a multidimensional measuring instrument is especially invaluable for African language lexicography, as the publication of paper dictionaries is still the default for these languages and it is precisely for paper dictionaries that these aspects are important.

His thorough knowledge of the grammar of Northern Sotho comes in handy when he ventures into the field of human language technology. He plays an important role in the development and evaluation of spell checkers for the African languages (De Schryver and Prinsloo 2004 and Prinsloo and De Schryver 2004). The development of tagsets and the annotation of text corpora is a logical consequence of his initial work in corpus compilation. In collaboration with local and international colleagues, he develops a tagset for Sepedi, as well as strategies for the automatic identification of nouns and verbs in this language. In 2015, Prinsloo and his two sons deliver a paper entitled *A Writing Tool for Sepedi* at the Elex Congress in the United Kingdom. The paper is very well received and attracts a lot of attention among the mainly international audience. The Sepedi Helper, as this tool is called, has been conceptualized as a support tool for dictionary users who need additional help with complex grammatical constructions. The ultimate goal of this research is the design and implementation of a complete electronic grammar for Sepedi that will eventually form part of an interactive e-dictionary. What distinguishes the Sepedi Helper from other, similar writing aid systems is the fact that it does not (only) function on the lexical level, but also on that of the morphosyntax. A user study, reported on in a follow-up article (Prinsloo and Taljard 2019), confirms that the Sepedi Helper does succeed in its goal: it confirms correct responses, it corrects errors and it identifies the most common errors that learners of Sepedi tend to make. Considering that Sepedi — like the other African languages of South Africa — is relatively poor with regard to electronic resources, the design and implementation of a functioning writing aid for this language is all the more impressive.

Together with renowned academic and fellow lexicographer Rufus Gouws, Prinsloo publishes a textbook *Principles and Practice of South African Lexicography* (Gouws and Prinsloo 2005), which is an indispensable reference work for any student in lexicography.

Prinsloo's status as an internationally renowned lexicographer is also reflected in the numerous chapters in highly regarded international publications, including *The Routledge Handbook of Lexicography* (Prinsloo, D.J., J.V. Prinsloo and Daniel Prinsloo 2018), *An International Encyclopedia of Lexicography* (Prinsloo 2013) and *eLexicography* (Prinsloo 2012b), to name but a few.

In addition to numerous papers read at international lexicography conferences including Asialex, EMLex, LREC and Australex, he is also invited in 2019 to deliver the keynote address at the 13th International Asialex Congress in Istanbul, Turkey, followed by a similar invitation to the 19th EURALEX Congress in Alexandroupoli, Greece.

As Alberts (this issue) points out, Danie Prinsloo was instrumental in founding Afrilex in 1995, with *Lexikos* as the organization's official magazine. It is therefore appropriate that a special edition of *Lexikos* be compiled to recognize the contribution that he made to lexicography.

Each of the contributions in this special issue represents an aspect of lexicography to which Danie Prinsloo contributed. However, he is not just a researcher; he is also a valued colleague, lecturer and supervisor. We hope that his retirement does not also imply his retirement from the world of lexicography, since his expertise is far too valuable for that.

This special edition would not have been possible without the support of the Bureau of the WAT and its staff. A special word of thanks therefore goes to Willem Botha, Tanja Harteveld and Hermien van der Westhuizen, who as usual, did an excellent job.

Bibliography

De Schryver, G.-M. and D.J. Prinsloo. 2000a. The Compilation of Electronic Corpora, with Special Reference to the African Languages. *Southern African Linguistics and Applied Language Studies* 18(1–4): 89-106.

De Schryver, G.-M. and D.J. Prinsloo. 2000b. Electronic Corpora as a Basis for the Compilation of African-language Dictionaries, Part 1: The Macrostructure. *South African Journal of African Languages* 20(4): 291-309.

De Schryver, G.-M. and D.J. Prinsloo. 2000c. Electronic Corpora as a Basis for the Compilation of African-language Dictionaries, Part 2: The Microstructure. *South African Journal of African Languages* 20(4): 310-330.

De Schryver, G.-M. and D.J. Prinsloo. 2000d. The Concept of 'Simultaneous Feedback': Towards a New Methodology for Compiling Dictionaries. *Lexikos* 10: 1-31.

De Schryver, G.-M. and D.J. Prinsloo. 2001. Fuzzy SF: Towards the Ultimate Customised Dictionary. *Studies in Lexicography* (11)1: 97-111.

De Schryver, G.-M. and D.J. Prinsloo. 2004. Spellcheckers for the South African languages, Part 1: The Status Quo and Options for Improvement. *South African Journal of African Languages* 24(1): 57-82.

Gouws, R.H. and D.J. Prinsloo. 2005. *Principles and Practice of South African Lexicography.* Stellenbosch: SUN PReSS, AFRICAN SUN MeDIA.

Prinsloo, D.J. 1979. *Lokatiefvorming in Noord-Sotho. Sotho. [Locative Formation in Northern Sotho.]* Unpublished M.A. thesis. Pretoria: University of Pretoria.

Prinsloo, D.J. 1984. *Woordvolgorde en volgordeverandering in Noord-Sotho. [Word Order and Order Change in Northern Sotho.]* Unpublished doctoral thesis. Pretoria: UNISA.

Prinsloo, D.J. 1987. Perspektief op pronominalisasie in Noord-Sotho. [Perspective on Pronominalization in Northern Sotho.] *South African Journal of African Languages* 7(1): 23-33.

Prinsloo, D.J. 1991. Towards Computer-assisted Word Frequency Studies in Northern Sotho. *South African Journal of African Languages* 11(2): 54-60.

Prinsloo, D.J. 1992. Lemmatization of Reflexives in Northern Sotho. *Lexikos* 2: 178-191.

Prinsloo, D.J. 2001. The Compilation of Electronic Dictionaries for the African Languages. *Lexikos* 11: 139-159.

Prinsloo, D.J. 2012a. Die leksikografiese bewerking van verwantskapsterme in Sepedi. [The Lexicographical Treatment of Kinship Terminology in Sepedi.] *Lexikos* 22: 272-289.

Prinsloo, D.J. 2012b. Electronic Lexicography for Lesser-resourced Languages: The South African Context. Granger, Sylvaine and Magali Paquot (Eds.). 2012. *Electronic Lexicography*: 119-144. Oxford: Oxford University Press.

Prinsloo, D.J. 2013. The Utilization of Bilingual Corpora for the Creation of Bilingual Dictionaries. Gouws, R.H. et al. 2013. *Dictionaries: An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography:* 1344-1356. HSK 5.4. Berlin/Boston: De Gruyter Mouton.

Prinsloo, D.J. and S.E. Bosch. 2012. Kinship Terminology in English–Zulu/Northern Sotho Dictionaries — A Challenge for the Bantu Lexicographer. Fjeld, Ruth Vatvedt and Julie Matilde Torjusen (Eds.). 2012. *Proceedings of the 15th Euralex International Congress, 7–11 August 2012, Oslo:* 296-303. Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo.

Prinsloo, D.J. and G.-M. de Schryver. 2002. Designing a Measurement Instrument for the Relative Length of Alphabetical Stretches in Dictionaries, with Special Reference to Afrikaans and English. Braasch, A. and C. Povlsen (Eds.). 2002. *Proceedings of the Tenth Euralex International Congress, EURALEX 2002, Copenhagen, Denmark, August 13–17, 2002*: 483-494. Copenhagen: Center for Sprogteknologi, Københavns Universitet.

Prinsloo, D.J. and G.-M. de Schryver. 2004. Spellcheckers for the South African Languages, Part 2: The Utilisation of Clusters of Circumfixes. *South African Journal of African Languages* 24(1): 83-94.

Prinsloo, D.J., J.V. Prinsloo and D. Prinsloo. 2018. African Lexicography in the Internet Era. Pedro A. Fuertes Olivera (Ed.). 2018. *The Routledge Handbook of Lexicography:* 487-502. London: Routledge.

Prinsloo, D.J., D. Prinsloo and J.V. Prinsloo. 2015. A Writing Tool for Sepedi. *E-Lex 2015*, Herstmonceux Castle, United Kingdom, 11–13 August 2015.

Prinsloo, D.J. and R.I. Raubenheimer. 1988. Troubleshooting Multiple-choice Items in Northern Sotho Item Banks for First Language Education. *South African Journal of African Languages* 8(3): 93-98.

Prinsloo, D.J. and E. Taljard. 2019. The Sepedi Helper Writing Assistant: A User Study. *Language Matters* 50(2): 73-99.

Raubenheimer, R.I. and D.J. Prinsloo. 1986. The Writing of Multiple-choice Items for Northern Sotho (First Language). *South African Journal of African Languages* 6 (Supplement): 101-135.

Raubenheimer, R.I. and D.J. Prinsloo. 1989. Item Analysis for Improving Multiple-choice Test Items in North Sotho. *South African Journal of African Languages* 9(2): 70-73.

Elsabé Taljard
Editor

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (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/10.5788/32-2-1705 (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/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1695 (Article)

Lexikos and AFRILEX — A Perfect Lexicographic Liaison

Mariëtta Alberts, *Former Director: Terminology and Standardisation, Pan-South African Language Board (PanSALB) (albertsmarietta@gmail.com)*

> *The essential point was the need for general recognition of the interdisciplinary nature of modern lexicographical work and the genuine will to cooperate.*
>
> (Reichling 1982)

Abstract: After in-depth discussions with interested parties in 1991 the Bureau of the Woordeboek van die Afrikaanse Taal (WAT) realized that the future of Afrikaans is inextricably connected with that of the other existing and utilized languages in South Africa. One of the results of the discussions was the establishment in 1991 of an academic journal *Lexikos* in the AFRILEX Series. An external feasibility study was also conducted on behalf of the Board of Control of the Bureau of the WAT to determine the possibility for the establishment of an Institute for Southern African Lexicography. The results of the feasibility study indicated that respondents did not want another bureaucratic institution. A major result of the feasibility study, however, was the establishment in 1995 of a professional association, the African Association for Lexicography, that concentrates exclusively on lexicographical issues. The Bureau of the WAT gave permission to the new association to use the acronym "AFRILEX". The Pan-South African Language Board (PanSALB), also established in 1995, was a direct consequence of the country's new multilingual dispensation. The legislation governing PanSALB was amended to allow for equal justice to all dictionary projects for the official South African languages. This led to the establishment of national lexicography units for each of the official South African languages. Both the activities of AFRILEX and the articles published in *Lexikos* have a huge influence on the activities of the national lexicography units.

Keywords: ASSOCIATION, COMMUNICATION, DICTIONARY, FEASIBILITY STUDY, JOURNAL, LANGUAGE FOR SPECIAL PURPOSES, LEGISLATION, LEXICOGRAPHY, NATIONAL LEXICOGRAPHY UNITS, TERMINOGRAPHY, TERMINOLOGY

Opsomming: *Lexikos* en AFRILEX — Perfekte leksikografiese samewerking. Na diepgaande samespreking met belanghebbendes in 1991 het die Buro van die Woordeboek van die Afrikaanse Taal (WAT) besef dat Afrikaans se toekoms onlosmaaklik verbind is aan die ander tale wat in Suid-Afrika bestaan en gebruik word. Een van die uitvloeisels van die samesprekings was die totstandkoming in 1991 van 'n vaktydskrif *Lexikos* in die AFRILEX-reeks. Die Beheerraad van die Buro van die WAT het ook 'n eksterne lewensvatbaarheidstudie laat doen om die moontlikheid vir die stigting van 'n Instituut vir Suider-Afrikaanse Leksikografie te bepaal. Die resultate van die lewensvatbaarheidstudie het getoon dat respondente nie nog 'n burokratiese

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1695 (Article)

instelling wou hê nie. 'n Verdere uitvloeisel van die lewensvatbaarheidstudie was egter die stigting in 1995 van 'n vakvereniging, die *African Assosciation for Lexicography*, wat uitsluitlik op leksikografiese aangeleenthede fokus. Die Buro van die WAT het toestemming gegee dat hierdie nuwe vakvereniging die akroniem "AFRILEX" gebruik. Die Pan-Suid-Afrikaanse Taalraad (PanSAT), wat ook in 1995 tot stand gekom het, was die direkte uitvloeisel van die nuwe meertalige beleid. Ten einde gelyke beregtiging vir alle woordeboekprojekte in al die amptelike Suid-Afrikaanse tale te weeg te bring, is die PanSAT-wetgewing gewysig. Dit het tot gevolg gehad dat nasionale leksikografiese eenhede vir elk van die amptelike Suid-Afrikaanse tale gestig is. Sowel AFRILEX se bedrywighede en die artikels wat in *Lexikos* gepubliseer word, het 'n groot invloed op die werksaamhede van die nasionale leksikografiese eenhede.

Sleutelwoorde: KOMMUNIKASIE, LEKSIKOGRAFIE, LEWENSVATBAARHEIDSTUDIE, NASIONALE LEKSIKOGRAFIESE EENHEDE, TERMINOGRAFIE, TERMINOLOGIE, VAKTAAL, VAKTYDSKRIF, VAKVERENIGING, WETGEWING, WOORDEBOEK

1. Introduction

Thirty years ago, in 1991, a perfect lexicographical liaison started with the publication of the first volume in an academic series of books and forged four years later in 1995 with the establishment of a professional association — both dedicated to the lexicography practice in their country of origin, South Africa. This perfect lexicographical liaison was formed between *Lexikos* in the AFRILEX Series and the African Association for Lexicography (AFRILEX) — both dedicated in their own unique way to enhance the principles, practice and usage of dictionaries in Africa and abroad and to assist with communication processes.

Language is the collective interest of the whole community and touches the communication needs of everyone in South Africa. The support for lexicographical activities and the compilation of various types of monolingual, bilingual and multilingual general dictionaries in the official languages and terminology lists in various subject areas underpins the South African policy of multilingualism, since it would contribute to the documentation of all official indigenous languages in the various languages or subject-related areas.

The Bureau of the Woordeboek van die Afrikaanse Taal (WAT) started publishing an academic journal *Lexikos* since 1991. The publication was one of the outcomes of a discussion held by the Board of Control of the Bureau of the WAT and various stakeholders. One of the most important aims with the publication of *Lexikos* in the AFRILEX series has always been to create a communication channel for national and international discussions on lexicographical issues, and in particular to serve lexicography in Africa, with its rich linguistic diversity.

As a result of an external feasibility study, conducted on behalf of the Bureau of the WAT the African Association for Lexicography (AFRILEX) was established in 1995. Lexicographic communication in Africa and abroad gained momentum with the establishment of AFRILEX and when *Lexikos* also officially

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1695 (Article)

became its mouthpiece. Officially *Lexikos* 6 was the first volume in the joint venture between *Lexikos* and AFRILEX.

It is important for any association to have a journal in which its members can publish the results of their research and their lexicographical endeavours. AFRILEX was in the privileged position to acquire in *Lexikos* a well-established journal with a scientific reputation respected both locally and internationally. The lexicographic discussion is stimulated during AFRILEX activities such as its annual international conferences and through the academic discourse in *Lexikos*.

In this article the focus is placed on the development of *Lexikos* and the establishment and activities of AFRILEX. The lasting liaison between *Lexikos* and AFRILEX and their influence on the establishment of the national lexicography units also receive attention.

2. *Lexikos*

Lexikos is a journal for the lexicographical specialist and enthusiast and is the only journal in Africa exclusively devoted to lexicography. The word "lexikos" is derived from the Greek word meaning "of or for words". *Lexikos* is since 1991 published by the Bureau of the Woordeboek van die Afrikaanse Taal in the AFRILEX Series. AFRILEX is the acronym for "lexicography in or for Africa".

2.1 Historical background

From 27 to 29 November 1989 formal discussions were held between the editorial staff of the Bureau of the Woordeboek van die Afrikaanse Taal and a group of prominent linguists on various lexicographical issues in order to transform the lexicographical practices of the Bureau. It was soon realized that there was an urgent need for the exchange of lexicographical and related knowledge and information, publications and experiences between South African lexicographers, linguists, academics, scholars as well as lexicographical and linguistic institutions abroad. A lexicographical growing point and stimulus was needed for the sharing of lexicographical activities and for lexicological reasoning in all South African languages. These discussions indirectly resulted in the establishment of *Lexikos*, which, as specialized academic journal, could serve as an excellent medium and vehicle for lexicographic discussions and exchange of information on lexicographical and terminographical research, principles and practice, experience, as well as management issues. It could contribute to, and stimulate academic discourse not only in Africa but also in the international lexicographical community. The AFRILEX series aimed to reflect on the implications of cooperation on the lexicographical profession and to fill the void of the lack of periodicals or monographs series dedicated solely to lexicography in Africa. It furthermore aimed to become a forum to discuss lexicography — a

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1695 (Article)

difficult but intellectually rewarding profession (Van Schalkwyk 1991: xx).

The first edition of *Lexikos* in the AFRILEX Series was published on 30 July 1991 by the Bureau of the Woordeboek van die Afrikaanse Taal (WAT). As Manager: Editorial Support Services at the Bureau of the WAT, Mr Pieter Harteveld was the first editor of *Lexikos*. The journal was fully prepared, typeset, proofread and financed by the Bureau. Ms Hermien van der Westhuizen, Ms Eleanor van Zuydam, and Ms Hanlie Meitzler of the Division Editorial Support Services assisted the editor with the editing and layout of the publication and the Bureau's Division Editorial Processing helped with proofreading. The cover of the journal was designed by Mr Piet Grobler.

Mr Etienne Botha, Ms Tanja Harteveld and Ms Riette Ruthven later joined the Division Editorial Support Services and provided assistance with the administration, editing, proofreading and electronic typesetting of the journal (Van Schalkwyk 1996: xiv). Ms Harteveld also acted/acts as review editor for recent publications.

2.2 Publications

The *Lexikos* journals were planned to be issues in a series of books (the AFRILEX Series) for specialists and all contributions should therefore be subject-related — dealing with pure lexicography on the one hand and linguistics, computer lexicography and management on the other hand (Harteveld 1991: xii). The first contributions published in *Lexikos* were only in Afrikaans and English. An English and an Afrikaans abstract containing an overview of the content of a contribution allowed for access to non-speakers of the language used in the specific contribution. In order to allow for discourse globally the journal later also included other languages such as Dutch, French and German. The articles should be accompanied by two summaries: one in English, French, German or Dutch, and the other in any other language used in Africa. In this way the main argument of the articles could be understood both locally and internationally (Du Plessis 2002: xi).

In the first issue of *Lexikos* (AFRILEX Series 1; 1991) the wish was expressed to promote lexicographical discussions across the whole of Africa by means of the publication. The second issue of *Lexikos* already indicated such improved cooperation with contributions from four countries outside South Africa, namely Egypt, Australia, Japan and the United States of America. The articles also showed a larger variety of languages which serve as objects of lexicographical research. In addition to Afrikaans (including Cape Muslim Afrikaans) and English, the South African languages Northern Sotho, Xhosa and Zulu, as well as Japanese, were in the second edition subjected to lexicographical investigation (Harteveld 1992: ix). Later volumes of *Lexikos* indicated global interest for lexicographical discourse and several other languages were lexicographically scrutinized.

Training is needed on both the formal and informal levels and in this regard *Lexikos* has a vital role to play. The emphasis of articles should ideally be on the theoretical and practical aspects of lexicography, but articles focusing on the training of dictionary compilers and on the education of dictionary users should also be included (Gouws 1998: xiv). Lexicographers and terminographers have to address the specific needs of the dictionary user. Dictionary compilers should know who their target users are and should familiarize themselves with the users' needs. Dictionaries should be user-friendly and users should be able to easily access and retrieve dictionary information. These are some of the burning issues experienced by *inter alia* the national lexicography units (NLUs) and which could be addressed by relevant articles in *Lexikos* (Alberts 2003: xiv-xv).

The focus *cum* scope of *Lexikos* is articles dealing with all aspects of lexicography and terminography or the implications that research in related fields of research such as linguistics, computer and information science may have for lexicography and all contributions in these fields are considered for publication. Initially only articles were included but later it became necessary to add other types of contributions and categories such as reviews and announcements. As from *Lexikos* 4 the contributions were distinguished typologically and classified under headings such as Articles, Review Articles; Projects; Reports and Lexiconotes. In recent years more headings were added such as: Research Articles; Contemplative Articles; On Learners' Dictionaries; Overview of Projects; Reviews; Lexiconotes; Lexicosoftware; Lexiconews; Lexicovaria; Lexicohonour; Lexicotribute; Lexicosurvey; Lexicobibliography, Lexicofocus; Prepublication Announcements; Publication Announcements; On the Compilation of Monolingual Dictionaries; Terminology Management; Meetings, and Corpora.

As from the first volume of *Lexikos* two or more highly-rated lexicographical experts adjudicated each article. Articles and review articles are subject to strict anonymous evaluation by independent academic peers in order to ensure the international research quality thereof. All types of contributions are peer reviewed and no concessions are made in this regard.

A *Lexikos* volume is published annually, but in 2009 *Lexikos* 19 and *Lexikos* 19 Supplement were published. The first 30 volumes have already been published (i.e. ISSN 2224-0039 [online]; ISSN 1684-4904 [print]). The initial volumes were published as hard copies and the publication of the printed issues was made possible *inter alia* by a generous donation from the L.W. Hiemstra Trust. As from 2018 *Lexikos* is published electronically and only a few hard copies were made available. Currently all editions are available online only. *Lexikos* has an open access policy in order to provide immediate open access to the content on the principle that making research freely available to the lexicography community and the public supports a greater global exchange of related knowledge. Since 2020 the journal follows a publish-as-you-go approach: as soon as an article has completed the review process and revisions (if any) had been made, it is immediately published online. *Lexikos* is available online at http://lexikos.journals.ac.za/ and from Sabinet, AJOL, Ebsco and Proquest.

From volume 4 onwards *Lexikos* has been accredited by the Department of

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1695 (Article)

Education as an income generating publication — this once again confirmed the high standard of the journal.

Lexikos holds a Creative Commons License CC BY 4.0. The journal is hosted by the Stellenbosch University Information Service (SU LIS) on request of the editor.

Lexikos is an indexed journal for which an Impact Factor is being calculated in the category Linguistics. With an Impact Factor of 0.667, *Lexikos* in 2009 rated 49th out of 92 journals in the Linguistics Category, which placed it in the so-called third quartile. In 2017 the Impact Factor of *Lexikos* as Linguistics Journal was higher than the average among journals in humanities published outside the Western World. This indicates that *Lexikos* established itself as one of the leading lexicographical journals in the world.

Lexikos is ISI-rated by Clarivate Web of Science on their various indexes, *inter alia* Arts and Humanities Citation Index®, Current Contents®/Arts & Humanities, Current Contents®/Social and Behavioral Sciences, Journal Citation Reports/Social Sciences Edition, Social Sciences Citation Index®, EuroPub Index and Social Scisearch®. *Lexikos* is furthermore indexed on Scopus, Linguistics Bibliography Online, Linguistics Abstracts Online, Linguistics and Language Behavior Abstracts, MLA International Bibliography and R.R.K. Hartmann's Bibliography of Lexicography.

2.3 Editors and management

With the publication of the international lexicographical journal *Lexikos* the Bureau of the Woordeboek van die Afrikaanse Taal shifted its focus to national and international lexicography. The establishment of *Lexikos* was one of the most important consequences of the Bureau's transformation into a modern dictionary office where dictionary compilation processes are conducted according to relevant metalexicographical principles (Botha 2003: xii). The first editors of *Lexikos* were all staff members of the Bureau.

The respective editors were/are responsible for the various editions of *Lexikos*. Since 1994 an Advisory Board as well as an Editorial Committee of international standing were appointed. The Advisory Board primarily gives advice to the editor and the Editorial Committee judges contributions for *Lexikos* to help ensure the academic quality of the publication.

The first editor of *Lexikos* was Mr Pieter Harteveld. He was the editor of the first five volumes of this publication but unfortunately suddenly passed away on 8 March 1996. His contribution to the establishment and expansion of *Lexikos* was remarkable. The scientific standard and relevance of the research contributions was of great importance to him. As Manager: Editorial Support Services at the Bureau of the WAT he saw to it that the presentation, layout and typography of *Lexikos* were easily accessible to the reader. His perfectionist attitude was apparent in the editorial treatment of every article in *Lexikos* (Van Schalkwyk 1996: xiii).

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1695 (Article)

After Mr Harteveld's passing Dr D.J. (Dirk) Van Schalkwyk, then Editor-in-Chief of the Bureau of the WAT, acted as editor of *Lexikos* 6 to finalize the publication. His ideal for cooperative lexicography that could lead to joint projects with tertiary institutions and dictionary units in Africa and Europe was reflected in this publication.

As from *Lexikos* 7 Dr J.C.M.D. (Johan) du Plessis took over as editor, that is from 1997 to 2010. When Dr Du Plessis retired from his position as the final editor at the Bureau of the WAT, he kept his post as editor of *Lexikos*, and he remained in this position until *Lexikos* 20 was published. He once again acted as editor of *Lexikos* 23 when the appointed editor was unavailable due to pressure of work and other commitments. During his editorship Dr Du Plessis established and strengthened the journal's position and status as scientific journal on an international level. He managed to create a true culture of scientific reporting on lexicography in Africa and abroad. He guided novice contributors by assisting them with their contributions.

Since 2011 *Lexikos* has been edited on a rotating triumvirate: Prof. Elsabé Taljard, Department of African Languages, University of Pretoria, Prof. D.J. (Danie) Prinsloo, Department of African Languages, University of Pretoria and Prof. R.H. (Rufus) Gouws, Department of Afrikaans and Dutch, Stellenbosch University. In 2016 Prof. Rufus Gouws had to withdraw from the team owing to other work and obligations. Dr Hugues (Steve) Ndinga-Koumba-Binza, University of the Western Cape and Département des Sciences du Langage, Université Omar Bongo, Libreville, Gabon then joined the editorial team. Due to the increased workload brought about by the editorship of Lexikos, Mr André du Plessis, co-editor at the WAT was also brought on board. Although the editorial team shares the joint responsibility for each volume, one member of the team is annually appointed as final editor. The current team of rotating editors is Prof. Elsabé Taljard, Prof. Dion Nkomo and Dr Hugues Steve Ndinga-Koumba-Binza, assisted by Mr Du Plessis.

Prof. Elsabé Taljard acted as editor of volumes 21, 25 and 28 of *Lexikos* and she is also the editor of *Lexikos* 31. With Prof. Taljard as editor at the helm of *Lexikos*, the future of this internationally recognized journal is bright. She is an expert in the fields of lexicography and terminography, especially those of the African languages and as editor she manages to cater for lexicography and terminography development on the African continent.

As editor of *Lexikos* 24 Prof. Rufus Gouws focused on an inclusive approach regarding the promotion of opportunities to discuss language-specific as well as general theoretical issues of lexicography. He furthermore stressed the value of *Lexikos* displaying a definite lexicographic and metalexicographic community of collaborators consisting of well-trained lexicographers, dictionary units and trained dictionary users. The contribution Prof. Gouws had made in establishing *Lexikos* as journal of choice for both national and international lexicographers can hardly be overestimated. His sound judgment and knowledge of both practical and theoretical lexicography have made him an outstanding

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1695 (Article)

editor (Taljard 2015: xii).

Prof. Danie Prinsloo, a founder member of AFRILEX, acted as the editor of *Lexikos* 22 in 2012. The publication again stimulated discourse between lexicographers as well as between lexicographers and linguists, and even laypeople. Collaborators continued to publish contributions to the lexicographic discussion in the specialist journal *Lexikos* in the AFRILEX Series, but monographic and other studies appeared as separate publications in the AFRILEX Series.

Prof. Prinsloo also acted as editor for *Lexikos* 26 in 2016 — the year when AFRILEX celebrated its 21st Birthday at the 21st International Conference of AFRILEX in Tzaneen, South Africa. This was befitting since he was instrumental in the establishment of AFRILEX. Prof. Prinsloo was at the time Head of the Department of African Languages at the University of Pretoria, and an outstanding scholar in the field of linguistics and lexicography, especially regarding research in African indigenous languages. His contributions towards *Lexikos* and AFRILEX contributed much to the development of lexicography and terminography on the African continent.

Prof. Elsabé Taljard and Dr Hugues Steve Ndinga-Koumba-Binza ably assisted Prof. Prinsloo with the compilation of *Lexikos* 26. *Lexikos* 29 was again compiled under the editorship of Prof. Danie Prinsloo and this time he was competently assisted by Prof. Dion Nkomo. Prof. Prinsloo not only contributed to the value of contributions published in *Lexikos* but he clearly also shares his experience in the field of theoretical and practical lexicography with other members by mentoring them also in aspects such as the editing of a prestigious academic publication such as *Lexikos*.

As co-editor of *Lexikos* 26, Dr Steve Ndinga-Koumba-Binza, regarded the 2016 volume of *Lexikos* under the editorship of Prof. Danie Prinsloo as a training phase. He also regarded his editorship of *Lexikos* 27 as a continuous learning experience, albeit enjoyable and manageable. There is, however, no doubt in the minds of AFRILEX members that he is capable of fulfilling his role as editor of *Lexikos*. He also acted as editor of *Lexikos* 30 in 2020.

2.4 *Lexikos* encouragement Prize for Scholarly Writing

While editor of *Lexikos,* Dr Johan du Plessis, in 2005, initiated the *Lexikos* encouragement prize for scholarly writing for entrants under 35 years of age. Contributions were awaited at the end of March of the relevant year and none of the entrants should previously have published more than two articles. This prize aimed to encourage students in lexicography and young lexicographers to conduct significant research in their field of study, and to raise the standard of scholarly writing in the field of lexicography. Competing for the prize therefore took the form of a scientific article. These articles could deal with a lexicographical or metalexicographical aspect or aspects of any language or languages used in Africa. Contenders to the prize submitted articles dealing with lexicographical or metalexicographical aspects of languages used in Africa, and

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1695 (Article)

the winning article was published in *Lexikos*. Although an amount of R1 500 was offered at the time, the real intention of the prize lied in its prestige value. In years when no candidates submitted articles, the prize was not presented. The Board, however, urged young lexicographers *cum* researchers to submit research articles.

2.5 *Lexikos* and AFRILEX

Since 1996 *Lexikos* is the official journal of the African Association for Lexicography (AFRILEX) (cf. paragraph 3). When it became the AFRILEX mouthpiece for discussions on lexicography and terminography, especially in Africa with its rich language diversity, one of the most important aims with the publication of *Lexikos* still remained, namely the establishment of a communication channel for national and international lexicographic discussion. *Lexikos* serves these fields *inter alia* by

— stimulating discussion between lexicographers, and between lexicographers and linguists;

— establishing and promoting contact with local and foreign lexicographic projects, to focus general attention to the intrinsic nature (principles and practice) of lexicography, which also involves fields such as lexicology, terminography, linguistics, general linguistics, computer and information science, management, etc.;

— fostering and coordinating cooperation in all fields of lexicography, and

— promoting the aims of AFRILEX (cf. ajol.info/index.php/lex).

3. The African Association for Lexicography (AFRILEX)

3.1 Background

Dictionaries play a vital role in communication and communication is vital for the well-being of all citizens of a country. Dictionary users may experience communication needs due to a lack of information in general and technical dictionaries and other lexicographical information resources (e.g. word or term banks). This could be related to insufficient dictionary products and/or lexicographical practice (Alberts 2005: 317). This statement was very true for South Africa during the bilingual political dispensation in the country (i.e. prior to 1994). The South African dictionary practice was fragmented. Different private, tertiary or government dictionary components or units were established as a result of historical, cultural, political or institutional reasons.

Only two dictionary units were funded by the South African government, i.e. the Bureau of the Woordeboek van die Afrikaanse Taal (WAT) in Stellenbosch and the Dictionary Unit for South African English (DSAE) in Grahams-

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1695 (Article)

town. This was as a result of the bilingual dispensation of the country at the time.

The "*Woordeboek van die Afrikaanse Taal*" Act (Act 50 of 1973) describes the activities and functions of the Bureau which was established in 1925. Several Parliamentary debates on the "*Woordeboek van die Afrikaanse Taal*" Amendment Act, 1991, further highlighted the position of the Bureau as lexicographic unit within the boundaries of South Africa. The Amendment Act was passed by Parliament in February 1991 and the substitutions and amendments provided greater autonomy to the Board of Control of the Bureau in regard to the appointment of staff and financial management (Bureau of the WAT 1991: 1, 9-10).

One of the tasks of the Bureau of the WAT is to liaise with South African, African, and overseas lexicographic and other linguistic institutions with a view to exchanging lexicographical knowledge, information and publications, and to act as a growth point and stimulus for lexicographical activities and lexicological thought (Bureau of the WAT 1991: 2). During the previous bilingual political dispensation (i.e. prior to 1994) advantageous liaison and collaboration was continued by the WAT with the Woordenboek der Nederlandsche Taal at the Instituut voor Nederlandse Lexicologie (INL) in the Netherlands, the Dictionary Society of North America (DSNA), the European Association for Lexicography (EURALEX), and the Dictionary Unit for South African English (DSAE) at the Rhodes University in Grahamstown. This was done by means of correspondence and the exchange of quarterly and annual reports (Bureau of the WAT 1991: 7-8). The Bureau, however, also felt the need for liaison with local dictionary projects compiled for the African languages.

In light of certain assumptions made on the position of Afrikaans in a future post-Apartheid political order and how this affected the financial positioning of the Bureau, the Board of Control and the Bureau started discussing the desirability and possibility of privatization of the Bureau in 1991 (Bureau of the WAT 1991: 5). The Board of Control furthermore discussed the possibility of establishing an Institute for Southern African Lexicography. It was decided to conduct a feasibility study to determine the attitudes, meanings and need for such an institute among interested parties within the lexicographical community in South Africa (Alberts 1993: xi). The then Editor-in-Chief of the Bureau of the WAT Dr D.J. (Dirk) van Schalkwyk requested an independent external research team consisting of Dr Mariëtta Alberts (project leader) and Prof. William Branford to conduct a feasibility study. The research was conducted during 1992 (Alberts 1993: 1, 4; Alberts 2005: 316-320).

3.2 Feasibility Study

In 1991 the Board of Control of the Bureau of the Woordeboek van die Afrikaanse Taal (WAT) requested a feasibility study to determine the need for a Southern African Institute for Lexicography (Alberts 1993). The purpose of such an Institute was to unite the different private, tertiary and governmental bodies that had been in place in the dictionary practice in South Africa as the outcome of historical, cultural and other reasons. While initiating the idea of an institute, the members of the Board of Control realized that such an institute could not be established without the consent, support and collaboration of all the stakeholders in the Southern African lexicography scene.

Funding for the feasibility study was obtained from GENCOR Development Trust. Although the Board of Control of the Bureau initiated the feasibility study the intended institute would not be financed by the Bureau nor any other South African dictionary unit. The institute would also not be managed nor controlled by the Board of Control of the Bureau of the WAT. The Bureau would share, like all other participants, an equal part in such an institute. The Board also realized that an institute could not be established without the collaboration and consent of all stakeholders in the Southern African lexicography fraternity (Alberts 1993: 1).

3.2.1 Problem statement

The premises that ruled the problem statement indicated that dictionaries play a significant role in providing proper communication and that proper communication was favourable for the welfare of every country and the speech communities. Dictionary users could experience communication needs as a result of poor information in dictionaries and other lexicographical information resources such as word and term banks due to an insignificant lexicographical practice (Alberts 1993: 1).

At the time of the research project (1992) still only the official dictionary offices of Afrikaans and English, i.e. the Bureau of the Woordeboek van die Afrikaanse Taal (WAT) and the Dictionary Unit for South African English (DSAE) respectively, received funding from the government of South Africa. One of the problems experienced by lexicographical units in South Africa was the lack of proper funding. The Bureau of the WAT and the English Dictionary Unit received government funding due to the official status of Afrikaans and English. The dictionary units of the local African languages did not share in this privilege to receive government funding — the few dictionaries compiled in these languages were either university-based or private initiatives, and therefore privately financed. Another problem was the lack of trained lexicographers and terminographers in the country (Alberts 2005: 317). Efficient communication and proper financial assistance were therefore needed to produce more dictionaries, to train lexicographers and terminologists for all Southern African languages and to provide proper general and technical dictionaries in all the languages used in South Africa (Alberts 1993: 1). It was furthermore thought that a national institute could disseminate available expertise and provide in-house training.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1695 (Article)

3.2.2 Research areas

Three research areas were determined for the feasibility study:

— the need for an Institute for Southern African Lexicography;

— the structure of such an institute;

— possible cooperation between various local stakeholders (Alberts 1993: 2).

3.2.3 Research methodology

The research was done according to proven research methodologies and practices (cf. Babbie 1983: 94; Alberts 1990: 17-20).

The targeted respondents received a covering letter, an information document explaining the purpose of the research, and a questionnaire. Prior to the formal study, the questionnaire was tested during a preliminary investigation among stakeholders (Alberts 1993: Annexure 6). Research information was obtained by means of the approved questionnaire that was available in English and Afrikaans and which contained 15 questions related to the research topic (Alberts 1993: Annexure 5).

3.2.4 Target group

The external needs assessment study targeted 186 known individuals involved in dictionary compilation, government-supported lexicographical projects, lexicography units, language offices, publishing houses, training institutions, publishers, educational or cultural committees of political groupings, lecturers teaching lexicography, terminography, linguistics, translation studies, heads of linguistics and language departments, lexicographers at universities, various opinion-formers, Ministers and MEC's responsible for education and/or cultural affairs, and language boards, cultural associations, political and cultural groupings, and all associations involved in linguistics and/or the lexicography practice at the time (Alberts 1993: Annexure 4).

3.2.5 Research findings

The respondents were very positive regarding future liaison possibilities. The main concern of the respondents turned out to be the lack of coordination of lexicographic efforts. The respondents wanted collaboration, training and the sharing of knowledge and expertise in the field, but they did not agree on a formal structure such as an institute. There was reservation about introducing a new formal bureaucratic controlling structure, underlined by the fear that an institute might hamper private initiatives and activities.

Some respondents suggested that a coordinating body, a clearinghouse or association, should be established as an interim structure before deciding on an institute. An association for lexicography could (re)unite interested parties, allaying underlying fears regarding individual projects and offer expertise, training, information, news, etc. rather than undermining any lexicographic efforts.

The response indicated that an Institute for Southern African Lexicography was viable but it also showed that the important role-players did not feel a need for such an Institute. The success of such an institute would have relied on their collaboration. The results of the feasibility study further indicated that respondents were afraid of a governing body with enforceable power, i.e. an authoritative and official body. They required a flexible structure with a lot of freedom. It was clear that there was at the time not a need for an official institute. It was strongly felt that whatever body is formed, it should not be bureaucratically structured and should not restrict individual freedom, *inter alia* with regard to management and control. A keen interest was, however, shown for a unifying body among lexicographers and members of related professions. There was also clearly a great need for collaboration, cooperation, coordination and communication — the four C's (Alberts 1993).

The response to the investigation into an Institute for Southern-African Lexicography (Alberts 1993: Annexures 7-30) could be summarized as follows:

— that the time was not ripe for the immediate establishment of a lexicographical institute;

— that a professional association for South African lexicography should be established as soon as possible to address the expressed needs of the respondents for communication between lexicographers, the study of lexicography, and a measure of cooperation between lexicographical activities and enterprises;

— that such an association could, if possible, function initially for the time being under the auspices of the Linguistic Society of Southern Africa (LSSA);

— that the new association should formulate its constitution and define its objectives and programme of action.

3.2.6 Recommendations

The study recommended that the professional association should, to ensure autonomy, have its own constitution and formulate its own aims and projects, to include:

— the establishment of a liaison office or clearing house to coordinate projects

— setting up an email network

— issuing a quarterly newsletter

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1695 (Article)

- publishing an accredited magazine (The possibility was mentioned that the Bureau of the WAT's already established journal *Lexikos* of the AFRILEX Series could fulfil this role.)
- organizing an annual conference to share professional information
- formulating a national policy regarding lexicography.

It was argued that:

- the new envisaged association would cost far less than a lexicographical institute. It would however, require fairly substantial funds, far in excess of what subscriptions from membership fees would yield, and its planners should address the question of funding immediately;
- the imbalance in the past between public expenditure on Afrikaans and English on the one hand and on African languages on the other, should be readdressed;
- lexicography depends to a great extent on publishing houses. Some publishers have an excellent record regarding the publication of dictionaries. Others try to make a profit without taking into consideration the overheads and research costs involved in the lexicographic compilation process. Publishers and newspaper corporations should be encouraged to effectively finance dictionary research and compilation, and to support the association financially if it comes into being.

Several suggestions were made on how to proceed:

- all respondents should be informed of the envisaged professional association;
- respondents who indicated interest in an association should be contacted to take part in the process of its planning;
- all the respondents and other interested parties would be invited to join the association and become members;
- a meeting should be called as soon as possible to gather stakeholders and interested parties for the establishment of such a professional association;
- feedback regarding the outcomes of the feasibility study should be given to all respondents;
- the report on the feasibility study should be made available to all stakeholders and decision-makers in the field of lexicography.

The research team therefore decided to suggest to the Board of Control of the Bureau of the WAT that the response on the feasibility study clearly indicated the need for a professional association for lexicography to be established to meet the needs of lexicographers and other related interest groups. The report

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1695 (Article)

on the feasibility study was presented to the Board of Control in 1992, and Afrikaans and English versions of the report were published in 1993 by the Bureau of the WAT. Members of the Board of Control were obviously not in full agreement with the conclusions of the study since they had hoped that the respondents would agree on the establishment of an Institute for Southern African Lexicography.

3.3 AFRILEX — then and now

3.3.1 Viability of an Association for Lexicography in Southern Africa

Apart from existing dictionary units and the national terminology office, the interest of lexicography and terminography in South Africa had up to the beginning of 1995 been served by the Linguistic Society of Southern Africa (LSSA) and the African Language Association of Southern Africa (ALASA). Lexicographers, however, according to the feasibility study felt the need for the establishment of an association dedicated to lexicography (Alberts 2005: 320).

In March 1995, Prof. D.J. (Danie) Prinsloo and Dr Mariëtta Alberts drafted a questionnaire to test the viability of an association for lexicography. Over 800 questionnaires were mailed to members of ALASA, LSSA, publishers, government departments, and even political parties.

The idea was greeted by overwhelming enthusiasm that left no option other than to establish an association. A postal nomination and voting procedure then followed. A Board was elected by postal vote with a ballot percentage of nearly 80%.

3.3.2 The establishment of AFRILEX in 1995

On 14 July 1995 several lexicographers, academics and stakeholders came together at the closure of the Eighth International Conference of the *African Languages Association for South Africa* (ALASA) to establish a professional lexicography association. At 11:00 Dr R.R.K. (Reinhard) Hartmann, who chaired the inaugural meeting and facilitated the whole process and the election of office bearers, officially announced the birth of the new member of the Lex family: the African Association for Lexicography.

Prof. R.H. (Rufus) Gouws was elected as Chairperson, Ms I.M. (Irene) Dippenaar and Prof. Sizwe Satyo were elected as Vice-Chairpersons, Dr Mariëtta Alberts as Secretary-Treasurer, Prof. Danie Prinsloo as Conference Organizer, and Mr Pieter Harteveld, the Editor of *Lexikos,* was coopted to the Board. Other Board members were: Prof. Adelia Carstens, Prof. Tony Links, Prof. Louis Louwrens, Prof. Buyiswa Mini, Prof. A.C. Nkabinde, Prof. Piet Swanepoel and Dr Dirk van Schalkwyk.

The Bureau of the WAT granted permission to this new Association to

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1695 (Article)

adopt the acronym AFRILEX as its name. After the establishment of the Association the Bureau's publication, *Lexikos*, became the official mouthpiece of the African Association for Lexicography (AFRILEX).

AFRILEX is fortunate to have an accredited magazine such as *Lexikos*, which is published by the Bureau of the WAT and serves to promote lexicography in its broadest sense. AFRILEX shares responsibility for the future existence of *Lexikos* with the Bureau of the WAT. AFRILEX membership fees do not cover the cost to produce *Lexikos*. The Bureau of the WAT attracts donations or sponsorships to be able to subsidize *Lexikos* publications.

AFRILEX publishes the journal *Lexikos* in the AFRILEX series and other appropriate literature. This creates an environment conducive to the exchange of ideas and to a mutual stimulus for researchers and practitioners in the fields of lexicography and terminography.

3.3.3 Management

AFRILEX is managed by a Board that is elected bi-annually at conferences or by postal or email ballot and holds annual general meetings during the annual international conferences. The first Board drafted a constitution that was adopted at the first Board meeting. Over the years some amendments were made to the constitution so as to keep it updated and relevant.

The Board and Executive work according to portfolios and each Board member has an allocated task to fulfil:

- President: oversees the activities of the association and its liaison with other associations;
- Vice-President: performs the responsibilities of the President when the President cannot do so;
- Secretary: keeps minutes, writes letters, and liaises with Board and general members;
- Registrar: updates the address lists, manages and maintains the AFRILEX database, and sends the address list to the editor and the compilers of *Lexikos* at the Bureau of the WAT;
- Treasurer: processes membership payments, changes the signatories, and prepares the auditor's report;
- Organizer: organizes seminars, workshops, tutorials and the annual conference, and liaises with other associations (e.g. to coordinate dates);
- Webmaster: maintains the website;
- Editor of *Lexikos*: serves on the Executive of the Board as *ex officio* member.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1695 (Article)

Each Board member tries to promote the Association at various conferences, seminars and symposia, e.g. by taking AFRILEX flyers to such gatherings. AFRILEX members also contribute to this drive (Alberts 2005: 322).

The following members served as Chairpersons/Presidents of AFRILEX:

— Prof. Rufus Gouws (1995–1998)

— Prof. Danie Prinsloo (1999–2002)

— Dr Mariëtta Alberts (2003–2006)

— Prof. Rufus Gouws (2007–2008)

— Prof. Gilles-Maurice de Schryver (2009–2012)

— Dr Maropeng Victor Mojela (2013–2016)

— Prof. Herman L Beyer (2017–2020)

The current Board (2021) consists of:

— Executive: Prof. Langa Khumalo (President); Prof. Sonja Bosch (Vice-President); Prof. Elsabé Taljard (Treasurer), Prof. Dion Nkomo (Secretary).

— Board Members: Dr Philip Louw, Dr Lorna Morris, Dr H. Steve Ndinga-Koumba-Binza and Mr André H. du Plessis.

The AFRILEX constitution makes provision for postal or email ballot as well as ballot at an annual general meeting (AGM). Initially postal nominations were received beforehand and the voting process was conducted at an AGM by paid-up members. Technically it is not a problem to vote at an AGM but members not attending an AGM do not have a chance to vote. AFRILEX decided upon a more democratic ballot system and maintains a bi-annual postal or email voting system. This system was decided upon because not all members are in a position to attend the annual general meeting. With a postal or email ballot, all members of the Association are able to bring out a vote.

3.3.4 Activities

AFRILEX aims to organize regular international conferences and seminars on topics relevant at a specific time. The Board coordinates conference dates with those of other local linguistic associations like ALASA and LSSA. A total of five continental associations for lexicography are currently active and AFRILEX cooperates with the other international lexicography associations, namely the European Association for Lexicography (EURALEX), the Asian Association for Lexicography (ASIALEX), the Australian Association for Lexicography (AUSTRALEX), the Dictionary Society of North America (DSNA), etc., and members attend conferences of these associations whenever possible. AFRILEX aims to coordinate conference dates with local and international associations (Alberts 2006).

AFRILEX also continually seeks cooperation with international terminological institutions such as TermNet, INFOTERM, ISO/TC 37 and its local counterpart SABS TC 37. Dr Mariëtta Alberts represented AFRILEX on INFOTERM and she also served as Vice-President of INFOTERM. She was a member of ISO/TC 37 and the convener of SABS TC 37. After her retirement Prof. Elsabé Taljard represented AFRILEX on these institutions. Prof. Dion Nkomo is AFRILEX's representative on Globalex.

AFRILEX holds its international conferences by invitation at tertiary or lexicographic institutions. The following annual International Conferences were held by AFRILEX:

— 1st International AFRILEX Conference (1996) Rand Afrikaans University, Johannesburg, South Africa

— 2nd International AFRILEX Conference (1997) University of Natal, Durban, South Africa

— 3rd International AFRILEX Conference (1998) Potchefstroom University for Christian Higher Education, Potchefstroom, South Africa

— 4th International AFRILEX Conference (1999) University of Pretoria, Pretoria, South Africa

— 5th International AFRILEX Conference (2000) University of Stellenbosch, Stellenbosch, South Africa

— 6th International AFRILEX Conference (2001) University of the North, Pietersburg, South Africa

— 7th International AFRILEX Conference (2002) Rhodes University, Grahamstown, South Africa

— 8th International AFRILEX Conference (2003) University of Namibia, Windhoek, Namibia

— 9th International AFRILEX Conference (2004) Omar Bongo University, Libreville, Gabon

— 10th International AFRILEX Conference (2005) The tenth anniversary of the association was celebrated at the 2005 conference hosted by *Sesiu Sesotho Dictionary Unit* at the University of the Free State, Bloemfontein, South Africa

— 11th International AFRILEX Conference (2006) University of Venda for Science and Technology, Thohoyandou, South Africa

— 12th International AFRILEX Conference (2007) Tshwane University of Technology, Pretoria, South Africa

— 13th International AFRILEX Conference (2008) Bureau of the Woordeboek van die Afrikaanse Taal, Stellenbosch, South Africa

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1695 (Article)

- 14th International AFRILEX Conference (2009) Xhosa Department, University of the Western Cape, Bellville, South Africa
- 15th International AFRILEX Conference (2010) University of Botswana, Gaborone, Botswana
- 16th International AFRILEX Conference (2011) University of Namibia, Windhoek, Namibia
- 17th International AFRILEX Conference (2012) University of Pretoria, Pretoria, South Africa
- 18th International AFRILEX Conference (2013) Nelson Mandela Metropolitan University, Port Elizabeth, South Africa
- 19th International AFRILEX Conference (2014) North-West University, Potchefstroom, South Africa
- 20th International AFRILEX Conference (2015) University of KwaZulu Natal, Durban, South Africa
- 21st International AFRILEX Conference (2016) Tzaneen, South Africa
- 22nd International AFRILEX Conference (2017) Rhodes University, Grahamstown, South Africa
- 23rd International AFRILEX Conference (2018) University of the Western Cape, Bellville, South Africa
- 24th International AFRILEX Conference (2019) University of Namibia, Windhoek, Namibia
- 25th International AFRILEX Conference (2020) that was scheduled to be held at the Stellenbosch University, Stellenbosch, South Africa was cancelled due to the global COVID-19 Pandemic
- 25th International AFRILEX Conference (2021) Not even a global pandemic deviates AFRILEX from its activities — the 25th International Conference was a fully virtual online conference due to the COVID-19 pandemic.

The structure of the international conferences over the years changed, but usually contains the following items on its agenda:

- Conference keynotes — one keynote speaker is usually an internationally renowned lexicographer or academic and the second keynote speaker is an expert from Africa in the field of lexicography
- Papers — these are adjudicated and abstracts of papers are collated in a Booklet of Abstracts
- Publishers' session — publishers are invited to discuss their latest publications and they can also exhibit their publications during conferences
- Software session — a demonstration of relevant software

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1695 (Article)

- Sessions on special projects and initiatives, e.g. SELA
- PanSALB session — discussion of work done by the national lexicography units (NLUs)
- Annual general meeting and bi-annual election of the Board and Executive
- Pre-conference workshop
- Post-conference excursion.

Members of AFRILEX are invited to share in lexicographic discussion through contributing to *Lexikos*. Over the years the editor of *Lexikos* has come to have the first choice for articles from reworked papers initially read at an AFRILEX conference. Additional contributions from any other author and from any other part of the globe on any language(s) are also published.

An AFRILEX Newsletter was disseminated between 1995 and 2009 (https://www.afrilex.co.za/publications) but unfortunately the editor of the newsletter usually had to write all the articles him-/herself. Some of the editors of the AFRILEX Newsletter were Dr Mariëtta Alberts, Dr Maropeng Mojela and Mr Motsamai Motsapi. In 2004/2005 there was a new effort to revive the AFRILEX Newsletter. As from 2009 onwards the newsletters were replaced by communication through circulars to the AFRILEX mailing list. AFRILEX manages a website and distributes lexicography related information online (https://www.afrilex.co.za). The AFRILEX website has been a successful instrument in promoting AFRILEX. There is also a very active WhatsApp group under the leadership of Dr Hugues Steve Ndinga-Koumba-Binza.

Membership of AFRILEX is open to all individuals who and institutions that have an interest in lexicography. The membership of AFRILEX is mostly comprised of dictionary compilers, members of the lexicography teams of the eleven NLUs, compilers of terminology lists or technical dictionaries for Language for Special Purposes (LSP), directors and members of various language boards and advisory bodies, lecturers and students of metalexicography and terminology, and other language practitioners such as translators, editors, interpreters, teachers and journalists. The members of AFRILEX have a responsibility towards the various speech communities they serve, helping to preserve the South African languages and develop them into functional languages in all spheres of life, in order for various language communities to enjoy and utilize their communication skills to the fullest.

3.4 Special recognition

The AFRILEX Board decided to recognise special contributions by members in honour of or on behalf of AFRILEX.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1695 (Article)

3.4.1 Certificate of merit

Certificates of merit in recognition of contribution to AFRILEX were presented to three deserving members so far: Dr Johan du Plessis, Editor of *Lexikos*, and two previous chairpersons, Prof. Rufus Gouws and Prof. Danie Prinsloo.

3.4.2 Honorary membership

The AFRILEX Constitution makes provision for bestowing honorary membership to an ambassador of AFRILEX and for lexicographical expertise. There are currently five honorary members, Prof. A.C. Nkabinde (July 2002), Prof. Rufus Gouws (July 2010), Dr Johan du Plessis (July 2012), Dr Mariëtta Alberts (July 2015) and Prof. Danie Prinsloo (2018). These honorary members all received a *laudatio* in honour of their contributions in the field of lexicography.

The Board decided in 2011 to change the AFRILEX Constitution regarding honorary membership by building a profile of the honorary members. It was decided that honorary members should not pay membership fees, that their names should be mentioned on letterheads in all official correspondence, and that their photographs should appear on the webpage.

3.4.3 AFRILEX pottery trophies

Dr Mariëtta Alberts, one of the founder members of AFRILEX and a keen potter, made pottery trophies of the AFRILEX emblem mounted on wood. These trophies were for several years awarded to keynote speakers and the conference organizer of International AFRILEX conferences. She also made a huge pottery AFRILEX emblem to be used by the Association during conferences or other events (Alberts 2005: 322).

3.5 Training

AFRILEX promotes and coordinates research in and the study and teaching of lexicography in Southern Africa in its broadest sense. AFRILEX therefore strives to be actively involved in all aspects of lexicography, whether practical or theoretical. Within the African context the need for lexicographic training is increasing. AFRILEX considers it as part of its responsibility to participate in training initiatives and encourages its members to become involved in training activities. AFRILEX aims to train lexicographers, terminologists and other language practitioners in various aspects relating to lexicographical and terminographical principles and practice.

AFRILEX undertook training initiatives such as SALEX '97 and AFRILEX-SALEX '98 and supports actions aimed at corpus creation and the development of computer programs. Locally the association also plays a major role in respect

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1695 (Article)

of dictionary compilation for the African languages and as a result AFRILEX is consulted on a regular basis and it supports individuals and organizations with advice and assistance (Gouws 1998: xiv; Prinsloo 1999: xv). Prof. Rufus Gouws played a leading role in several PanSALB-initiated lexicographic training opportunities and in 2001 was the co-presenter of a two-day seminar on the compilation of dictionaries for special purposes (LSP) (Prinsloo 2001: xv).

Since the establishment of national lexicography units Prof. Rufus Gouws, Prof. Danie Prinsloo and Dr Mariëtta Alberts were involved in the training of the editors-in-chief and their staff at the respective NLUs. The Editors-in-Chief of the NLUs as well as members of PanSALB attended training sessions at the Bureau of the WAT on aspects such as the administration and management of NLUs.

Prof. Gilles-Maurice de Schryver, another very active member of AFRILEX, executed groundbreaking work both in the compilation of electronic corpora for the African languages and through adapting the *Onoma* computer program for the compilation of African language dictionaries (Prinsloo 2001: xv). Prof. De Schryver took his membership of AFRILEX very seriously by even marrying a fellow AFRILEX member, Ms Minah Nabirye in August 2009.

AFRILEX aims to collaborate with as many lexicographic and terminographical bodies and associations as possible because such cooperation is to the benefit of its members. The 6th International TAMA Conference, co-hosted by AFRILEX was held in South Africa on 17–21 February 2003. TAMA donated its surplus funds, accumulated for the event, to AFRILEX to be used for the future presentation of terminology or lexicography workshops.

AFRILEX participates in various training sessions. One such training session it participated in was *TermTrain*. *TermTrain* is a project within the framework of the *UNESCO Information for All Programme*. In September 2005 a *TermTrain* workshop was held in Benoni.

A follow-up *TermTrain* workshop, namely *TermTrain*: *Train the Trainer* was held from 27 to 31 March 2006 at the SABS in Pretoria. This training workshop was organized by TermNet (Austria) in collaboration with Standards South Africa (SABS), PanSALB, the National Language Service, DAC (currently Department of Sport, Arts and Culture (DSAC)), and AFRILEX. Trainees from all over South Africa attended this training workshop. The trainers were terminology and IT experts from Germany, Belgium, Austria and South Africa. Prof. Rufus Gouws and Dr Mariëtta Alberts were two of the trainers, while Mr David Joffe presented a demonstration of TshwaneTerm. Prof. Elsabé Taljard was the other member of the Executive that represented AFRILEX. It was an extremely specialized training session and the delegates indicated that they had gained a lot from attending (Alberts 2006).

On 14 August 2010 an AFRILEX workshop, the first in a series, was held on Dictionary Use in Mamelodi on the Mamelodi campus. The TAMA surplus was used for this workshop since it was within the boundaries set by TAMA for these funds. This workshop was designed as pilot to present to the Department of Education to conduct this kind of training on a more formal way at other schools. The presenters were Prof. Danie Prinsloo and Dr Victor Mojela, who trained 80 primary and 60 secondary school learners on dictionary usage through medium of Northern Sotho. Prof. Rufus Gouws trained 13 teachers on the principles and practice of dictionary use. Copies of MML Foundation Phase dictionaries were given to the primary school learners and the secondary school learners received in addition to the MML dictionaries also copies of the Oxford Northern Sotho dictionary. The very successful workshop empowered the learners and teachers and more such future workshops were also planned for learners in the Western Cape.

AFRILEX members also attended the LaRC Conference: "Controlled Natural Language" on 8 June 2013, as well as the ISO/TC 37 plenary sessions held on 9–14 July 2013 in Pretoria.

Prof. Pedro A. Fuertes-Oliviera of the University of Valladolid, Spain conducted a successful AFRILEX workshop on specialized lexicography.

3.6 Grants, prizes and sponsorships

3.6.1 Kernerman Grants

Over several years, AFRILEX members were fortunate to receive Kernerman Dictionary Research Grants. These grants totalling $1000 enabled the recipients to continue with research projects in lexicographical matters or to complete masters or doctoral studies in lexicography. Kernerman Dictionaries later decided not to continue with the presenting of grants. There was, however, a continued working relation between AFRILEX and Kernerman Dictionaries. In 2006 Ian Kernerman published an article on the establishment of AFRILEX, written by Dr Mariëtta Alberts, in the *Kernerman Newsletter June 2006* (Prinsloo 2002: xv; Alberts 2005: 323; Alberts 2006).

3.6.2 Other grants

In the past AFRILEX members also successfully applied for grants from EURALEX, ASIALEX, AUSTRALEX and the DSNA.

3.6.3 Laurence Urdang Award for Lexicography

AFRILEX members, who are also a member of either EURALEX or the DSNA could apply for the Laurence Urdang Award for Lexicography (Alberts 2006).

3.6.4 Publishers

Since the establishment of AFRILEX, Pharos Publishers, one of the leading local

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1695 (Article)

publishers of dictionaries, sponsored or co-sponsored the AFRILEX Conference dinner. This was soon regarded as a Conference tradition and was much appreciated by the members of the Association (Alberts 2005: 323). Pharos also sponsored conference bags.

The Conference Dinner of 2005 was again partially sponsored by Pharos Publishers. Pharos, however, decided not to continue with this tradition but to rather present a grant to assist members in attending Conferences.

Other publishers, such as Oxford University Press (OUP), offer discounts to AFRILEX members on their products.

3.7 The way forward

None of the members of AFRILEX are impartial towards lexicography and they are all interested to learn more about lexicography, dictionaries and how to compile dictionaries or use them to their best advantage. Members of AFRILEX are all in one way or the other involved in dictionary compilation, in different kinds of dictionary work (general dictionaries or dictionaries for Language for Special Purposes [LSP]) or projects, some are members of the lexicography teams of the eleven national lexicography units (NLUs), some are in an advisory capacity (e.g. PanSALB, its subcommittees, and structures such as NLBs and PLCs), lecturers in lexicography and terminography at tertiary institutions, students in these fields, language practitioners such as translators, editors, interpreters, language teachers, journalists, and publishers, or for the mere love for or addiction to lexicography (Alberts 2005: 323- 324).

It is clear that AFRILEX has, since its establishment, supported cooperative lexicography and terminology through its international conferences, symposia, various lexicography related activities and the publication of excellent articles in the *Lexikos* journal.

The members of AFRILEX have a responsibility towards the various speech communities they serve. A dictionary culture needs to be created and speech communities should be aware of AFRILEX. AFRILEX will have a bright future if all members of AFRILEX would continue to further the Association's aims in the Southern region of Africa, Africa itself and globally.

4. Dictionaries and Education

In October 2010 the Department of Higher Education and Training held a Roundtable discussion on the position and developmental status of the African languages. According to the Language Policy in Higher Education, the use of South African languages in instruction and learning in higher education will require development of dictionaries and other teaching and learning materials. The existence of various general dictionaries compiled by the National Lexicography Units of PanSALB; the various multilingual terminology lists compiled

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1695 (Article)

by the terminologists of the Terminology Coordination Section (TCS) of the National Language Service (NLS), Department of Arts and Culture (DAC); and other teaching and learning materials that have been produced were acknowledged. The questions, however, raised by the delegates were whether these dictionaries and materials were enough to meet the requirements of the target group and whether more should be done to meet the requirements of constitutional multilingualism and target users. It was agreed that the State should start with systematic and widespread translation into African languages, supported by lexicography work and terminology development. It was further agreed that more funding should be made available to dictionary compilation efforts, albeit general of technical dictionaries (Alberts, Botha and Kapp 2010; Nosilela 2010). In 1991 Bamgbose (1991: 109-111) already mentioned that there is an escalating consciousness in Africa and worldwide on the positive impact of multilingualism, especially on the role of African Languages in advancing multilingualism in education. This is reflected in the language policies that acknowledge the need for the teaching in, and acquisition of these languages at all levels of education.

5. Conclusion

Several contributing factors had an influence on the South African lexicography practice since 1991. The Bureau of the Woordeboek van die Afrikaanse Taal decided to restructure, and one of the results was the first publication of a professional journal *Lexikos* in the AFRILEX Series. The Bureau of the WAT also at the time initiated the idea for the establishment of an Institute for Southern African Lexicography. An independent research team was appointed to conduct the research. The external feasibility study of 1992 indicated that stakeholders did not want another bureaucratic institution and rather suggested the establishment of a clearing house or association to realize the need for collaboration, cooperation, coordination and communication in the field of lexicography. Further research by Prof. Danie Prinsloo and Dr Mariëtta Alberts indicated the explicit need for an association and in 1995 the African Association for Lexicography (AFRILEX) was established. *Lexikos* remains the mouthpiece of AFRILEX and all its members and continue to play a significant role in disseminating lexicographical information in this country, Africa and globally. It is clear that all these institutions form a perfect lexicographic liaison.

6. Bibliography

Alberts, M. 1990. *'n Bepaling van Afrikaanse vakleksikografiese behoeftes.* D.Litt. et Phil. Thesis. Pretoria: Unisa.

Alberts, M. 1993. *Feasibility Study: Institute for Southern African Lexicography/Lewensvatbaarheidstudie: Instituut vir Suider-Afrikaanse Leksikografie.* Stellenbosch: Bureau of the WAT.

Alberts, M. 2003. A Few Words from AFRILEX. *Lexikos* 13: xiv-xv.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1695 (Article)

Alberts, M. 2005**.** The African Association for Lexicography: After Ten Years. *Lexikos* 15: 316-324.

Alberts, M. 2006. *Standardisation, Modernisation and Harmonisation — User's Perspectives*. Paper read at the 11th International Afrilex Conference, 6–7 July 2006, Thohoyandou.

Alberts, M., W.F. Botha and P.H. Kapp. 2010. *Historical Experiences in Developing Afrikaans as a Language.* Paper presented at the Roundtable on African Languages, organized by the Department of Higher Education and Training, held on 22 October 2010, at Unisa, Sunnyside Campus, Pretoria.

Babbie, E. 1983. *The Practice of Social Research*. Belmont, California: Wadsworth.

Bamgbose, A. 1991. *Language and the Nation: The Language Question in the Sub-Saharan Africa.* Edinburgh: Edinburgh University Press.

Botha, W. (Ed.). 2003. *'n Man wat beur. Huldigingsbundel vir Dirk van Schalkwyk.* Stellenbosch: Bureau of the WAT.

Bureau of the Woordeboek van die Afrikaanse Taal. 1991. *Report of the Board of the Woordeboek van die Afrikaanse Taal for the period 1 April 1990 to 31 March 1991.* Stellenbosch: Bureau of the WAT.

Du Plessis, J.C.M.D. 2002. Foreword. *Lexikos* 12: xi-xii.

Goetschalckx, J. and L. Rolling (Eds.). 1982. *Lexicography in the Electronic Age*: *Proceedings of a Symposium held in Luxembourg, 7–9 July 1981.* Amsterdam: North-Holland Publishing Company.

Gouws, R.H. 1998. A Few Words from AFRILEX. *Lexikos* 8: xiv.

Harteveld, P. 1991. Foreword. *Lexikos* 1: xii-xiv.

Harteveld, P. 1992. Foreword. *Lexikos* 2: ix-x.

Nosilela, B. 2010. *Current State of African Languages in Institutions of Higher Learning.* Paper presented at the Roundtable on African Languages, organized by the Department of Higher Education and Training, held on 22 October 2010, at Unisa, Sunnyside Campus, Pretoria.

Prinsloo, D.J. 1999. A Few Words from AFRILEX. *Lexikos* 9: xv.

Prinsloo, D.J. 2001. A Few Words from AFRILEX. *Lexikos* 11: xv.

Prinsloo, D.J. 2002. A Few Words from AFRILEX. *Lexikos* 12: xv-xvi.

Reichling, A. 1982. Summary of the Round Table Discussion. Goetschalckx, J. and L. Rolling (Eds.). 1982: 267.

Van Schalkwyk, D.J. 1991. Words Require Contact. *Lexikos* 1 xviii-xx.

Van Schalkwyk, D.J. 1996. Foreword. *Lexikos* 6: xiii-xiv.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

Optimalisering van gratis elektroniese/aanlyn hulpbronne vir woordeboeksamestelling — 'n drietalige woordeboekeksperiment

Sonja E. Bosch, Departement Afrikatale,
Universiteit van Suid-Afrika, Pretoria, Suid-Afrika
(*seb@hbosch.com*)
ORCID: *https://orcid.org/0000-0002-9800-5971*

Marissa Griesel, Departement Afrikatale,
Universiteit van Suid-Afrika, Pretoria, Suid-Afrika
(*griesm@unisa.ac.za*)
ORCID: *https://orcid.org/0000-0003-1309-0212*
en
Elsabé Taljard, Departement Afrikatale,
Universiteit van Pretoria, Pretoria, Suid-Afrika
(*elsabe.taljard@up.ac.za*)
ORCID: *https://orcid.org/0000-0002-4507-1633*

Opsomming: Die beskikbaarheid van meertalige woordeboeke is deurslaggewend, nie slegs vir direkte teikengebruikers nie, maar ook vir indirekte teikengebruikers soos menslike taaltegnoloë, veral in die geval van tale met skaars hulpbronne, soos byvoorbeeld Venda. In hierdie artikel word die optimale benutting van gratis elektroniese/aanlyn hulpbronne vir die samestelling van 'n drietalige e-woordeboek vir Venda, Engels en Afrikaans ondersoek. Ons benadering is gebaseer op 'n eksperiment waarin die samestellingsproses so ver moontlik geoutomatiseer is om besparing in terme van tyd en mensekrag teweeg te bring. Engels word as 'n brug vir die vertaling tussen die brontaal, Venda, en die doeltaal, Afrikaans, gebruik. Die algemene bevindinge is dat daar sekere beperkings te wagte kan wees in so 'n semi-outomatiese proses wat wel 'n sekere mate van menslike intervensie verg. Hoewel die saamgestelde e-woordeboek nie as 'n finale produk beskou kan word nie, bied die woordeboeksamestellingsprogram *Lexonomy*, wat vanweë sy aanpasbaarheid en maklike uitleg suksesvol in hierdie studie gebruik is, die geleentheid vir menslike insette om die nodige aanpassings op 'n gebruikersvriendelike wyse te doen. Die geformuleerde konsepvoorstel is nuttig vir die skep van meertalige aanlyn woordeboeke, saamgestel met behulp van beskikbare aanlyn of elektroniese hulpbronne. Die resulterende drietalige woordeboek is aanlyn beskikbaar as bewys van die konsep waarop verdere werk kan bou. Die feit dat die databasis onderliggend aan die woordeboek beskikbaar is in 'n masjienleesbare formaat, naamlik XML, is belangrik vir indirekte

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

teikengebruikers vir hergebruik om elektroniese hulpbronne te ontwikkel, veral vir hulpbronarm tale.

Sleutelwoorde: WOORDEBOEKSAMESTELLING, VENDA–ENGELS–AFRIKAANS, DRIETALIGE WOORDEBOEK, ELEKTRONIESE/AANLYN HULPBRONNE, MASJIENVERTAALSISTEME, *LEXONOMY*, KORPUSSOEKTOG, TEIKENGEBRUIKERS

Abstract: Optimization of Free Online/Electronic Resources for Dictionary Compilation — A Trilingual Dictionary Experiment. The availability of multilingual dictionaries is crucial, not only for direct target users, but also for indirect target users, especially in the case of languages with scarce resources such as Venda. This article explores the optimal use of free electronic/online resources for compiling a trilingual e-dictionary for Venda, English and Afrikaans. Our approach is based on an experiment in which the compilation process was automated as far as possible to achieve savings in terms of time and manpower. English is used as a bridge for the translation between the source language, Venda, and the target language, Afrikaans. The general finding is that certain limitations can be expected in such a semi-automated process that requires a certain amount of human intervention. Although the composite e-dictionary cannot be considered a final product, the dictionary compilation program *Lexonomy*, which has been used successfully in this study due to its adaptability and easy layout, provides the opportunity for human input to make the necessary adaptations in a user-friendly manner. The proposed concept is useful for creating multilingual online dictionaries, compiled using available online or electronic resources. The resulting trilingual dictionary is available online as proof of concept on which further work can build. The fact that the database underlying the dictionary is available in a machine-readable format, namely XML, is important for indirect target users for reuse to develop electronic resources, especially for resource-scarce languages.

Keywords: DICTIONARY COMPILATION, VENDA–ENGLISH–AFRIKAANS, TRILINGUAL DICTIONARY, ELECTRONIC/ONLINE RESOURCES, MACHINE TRANSLATION SYSTEMS, *LEXONOMY*, CORPUS SEARCH, TARGET USERS

1. Inleiding

Woordeboeksamestelling vir tale met skaars hulpbronne is 'n arbeidsintensiewe taak. In hierdie artikel word die optimale benutting van gratis elektroniese/ aanlyn hulpbronne vir die saamstel van 'n drietalige woordeboek ondersoek en bespreek. Die betrokke tale is Venda, Engels en Afrikaans. Die doelstellings van hierdie studie is eerstens om met 'n unieke benadering te eksperimenteer waar Engels as 'n brug vir die (semi-outomatiese) vertaling tussen die brontaal, Venda en die doeltaal, Afrikaans, gebruik word. 'n Verdere doelstelling is die formulering van 'n konsepvoorstel wat gebruik kan word vir die daarstel van 'n drietalige aanlyn woordeboek, saamgestel met behulp van gratis hulpbronne. Ons ondersoek in die derde plek die mate waartoe die samestellingsproses geoutomatiseer kan word, aangesien 'n (semi-) outomatiese benadering 'n besparing in terme van tyd en mensekrag teweeg kan bring. Die semi-outomatiese benadering wat in hierdie studie voorgestel en beskryf word, verhoog ook die toeganklikheid tussen minderheidstale (Venda en Afrikaans). Die resulterende drietalige woordeboek sal verder met behulp van gratis hulpbronne ook aanlyn beskikbaar gestel word as 'n bewys van die konsep waarop verdere werk kan bou.

In die volgende afdeling word 'n oorsig gegee van bestaande woordeboeke met Venda en Afrikaans as taalpaar, waarna twee tipes potensiële teikengebruikers in Afdeling 3 geïdentifiseer word. Afdeling 4 beskryf die beskikbare gratis aanlyn/elektroniese hulpbronne waarop die eksperiment (soos beskryf in Afdeling 5) gebaseer is. Dit word gevolg deur 'n evaluering van die eksperiment in Afdeling 6 en laastens, in Afdeling 7, voorstelle vir toekomstige werk.

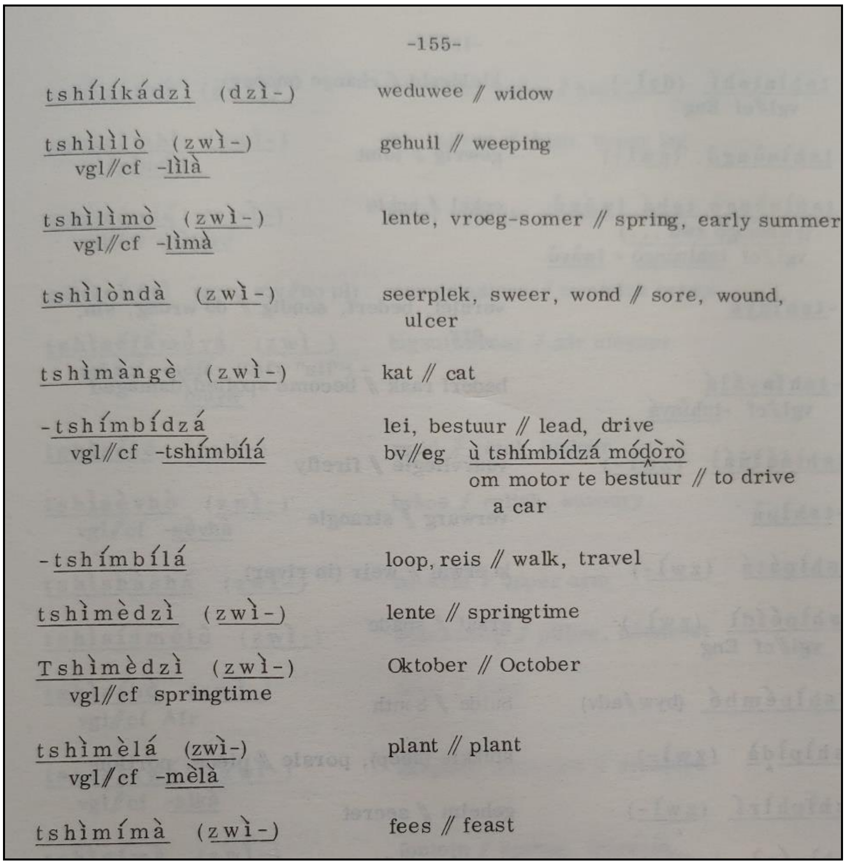
2. Bestaande woordeboeke met Venda en Afrikaans as taalpaar

Die enigste woordeboek waarin Afrikaans en Venda as taalpaar voorkom, is die *Drietalige Elementêre Woordeboek* (DEW) deur Wentzel en Muloiwa (1976). In die voorwerk (ibid: i) van die woordeboek word die doel van die woordeboek soos volg verwoord:

> Hierdie elementêre drietalige woordeboek is bedoel om — al is dit totdat 'n meer volledige uitgawe kan volg — in 'n dringende behoefte te voorsien. Daar is naamlik geen Venda-woordeboek beskikbaar nie en dit is bykans onmoontlik om byvoorbeeld die Vendataal te probeer aanleer sonder 'n beskikbare woordeboek .... Die enigste woordeboek wat nog ooit saamgestel is, is die *Tshivenḓa–English Dictionary* deur NJ van Warmelo (1937). Hierdie boek is egter reeds vir baie jare uit druk en hoewel dit 'n baie betroubare naslaanwerk is, is dit nie vryelik beskikbaar vir studente nie.

Die DEW bestaan uit drie dele. In die eerste deel is die taal van lemmatisering Venda, met Afrikaanse en Engelse ekwivalente, deel twee bestaan uit 'n Afrikaans–Venda-afdeling, gevolg deur deel drie, die Engels–Venda-afdeling. Die woordeboek het 'n raamstruktuur bestaande uit 'n voorwerk, gevolg deur die drie alfabetiese lemmalyste. Die lemmatiseringstrategie in deel een is woordgebonde, soos meestal die geval is vir disjunktiefgeskrewe Afrikatale. Die ordening is nie streng alfabeties nie — waar lemmas met konsonante begin wat 'n diakritiese teken gebruik om die Romeinse alfabet aan te vul, word hierdie lemmas in 'n aparte alfabetiese strek geplaas. Die dentale simbole ḓ, ḽ, ṋ en ṱ gaan die gewone d, l, n en t vooraf, terwyl die velêre ṅ ná gewone n volg. Die artikelstruktuur is relatief eenvoudig. Woordkategorieë word slegs op implisiete wyse aangedui — in die geval van naamwoorde word die meervoud aangedui deur die meervoudsprefiks in hakies ná die lemma te verstrek, gevolg deur die Afrikaanse en Engelse vertaalekwivalente. Vergelyk byvoorbeeld die lemma *tshimedzi* 'lente' in Figuur 1:

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)



-155-

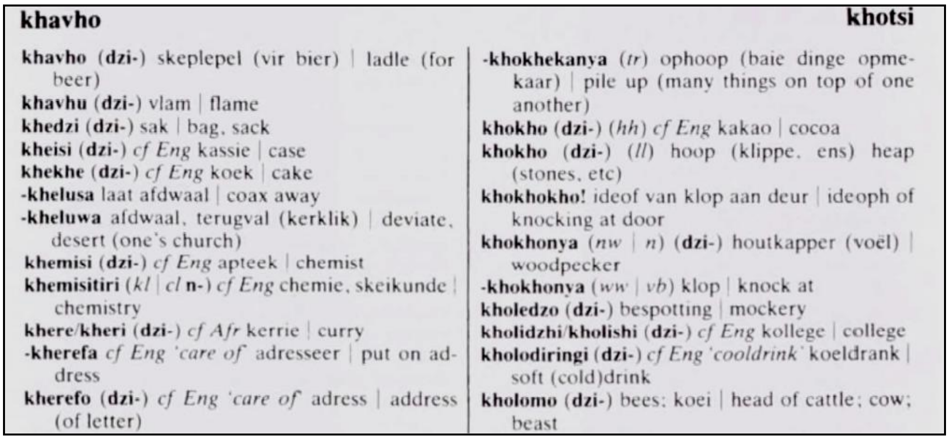
tshílíkádzì (dzì-)	weduwee // widow
tshìlìlò (zwì-) vgl//cf -lìlà	gehuil // weeping
tshìlìmò (zwì-) vgl//cf -lìmà	lente, vroeg-somer // spring, early summer
tshìlòndà (zwì-)	seerplek, sweer, wond // sore, wound, ulcer
tshìmàngè (zwì-)	kat // cat
-tshímbídzá vgl//cf -tshímbílá	lei, bestuur // lead, drive bv//eg ù tshímbídzá móḓòrò om motor te bestuur // to drive a car
-tshímbílá	loop, reis // walk, travel
tshìmèdzì (zwì-)	lente // springtime
Tshìmèdzì (zwì-) vgl//cf springtime	Oktober // October
tshìmèlá (zwì-) vgl//cf -mèlà	plant // plant
tshìmímà (zwì-)	fees // feast

Figuur 1: Uittreksel uit die *Drietalige Elementêre Woordeboek* deur Wentzel en Muloiwa (1976: 155)

In gevalle waar meervoudsvorming onreëlmatig is of waar fonologiese verandering plaasvind, word die volledige meervoudsvorm verstrek — vergelyk byvoorbeeld *danda* 'paal', meervoud *matanda*. In die geval van deverbatiewe naamwoorde word 'n kruisverwysing na die werkwoordstam wat as basis vir die deverbatief dien, aangegee, sien byvoorbeeld *tshimela* 'plant (naamwoord)' met kruisverwysing na *-mela* 'plant (werkwoordstam)' in Figuur 1 hierbo. By werkwoorde as lemmas word die werkwoordstatus aangedui deur 'n koppelteken wat die stam voorafgaan, vergelyk *-tshimbila* 'loop'. Afgeleide werkwoordstamme word gelemmatiseer en behandel, maar 'n kruisverwysing na die basisvorm word ook verstrek, byvoorbeeld *-tshimbidza* 'lei, bestuur' met kruisverwysing na *-tshimbila* 'loop'. Toon word in alle gevalle aangedui. Etimologiese inligting word in die geval van leenwoorde uit Afrikaans of Engels verskaf, en gebruiksvoorbeelde word op skynbaar lukrake wyse verstrek.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

In 1982 verskyn 'n verbeterde uitgawe van hierdie woordeboek wat in 2009 'n sesde druk beleef. Waar die 1976-weergawe op studente gemik was wat die toenmalige Spesiale Kursus in Venda aan UNISA gevolg het, word daar in die voorwoord tot die 1982-weergawe aangedui dat 'n breër spektrum gebruikers die teiken van dié uitgawe is. Dit word beplan as die eerste in 'n reeks van drie woordeboeke: 'n praktiese klein woordeboek, wat beplanningsgewys gevolg sou word deur 'n omvattende twee- of drietalige woordeboek wat onder andere ook idiomatiese taalgebruik en die aanduiding van toon sal insluit, en 'n naslaanwerk soortgelyk aan dié van N.J. van Warmelo se *Tshivenḓa–English Dictionary* (1937). In die 1982-weergawe word die aantal inskrywings verdubbel, maar voorbeelde word beperk tot enkele gevalle waar dit as onontbeerlik beskou word. Daar word ook weggedoen met die aanduiding van toon, behalwe in gevalle waar toon 'n woordonderskeidende funksie het, vergelyk byvoorbeeld *khokho* 'kakao' en *khokho* 'hoop klippe' in Figuur 2 hieronder. Toon word ook nie ortografies aangedui nie, maar met behulp van die letters *h* = hoog en *l* = laag. Kruisverwysings na basisvorms in geval van afgeleide werkwoordstamme en deverbatiewe word ook weggelaat. Enkele gebruiksnotas word in hakies verstrek, bv. "*evho*! O nee! (slegs vroue)" (Wentzel en Muloiwa 1982: 14). Vergelyk Figuur 2 vir 'n uittreksel uit die verbeterde weergawe:



khavho **khotsi**

khavho (**dzi-**) skeplepel (vir bier) | ladle (for beer)
khavhu (**dzi-**) vlam | flame
khedzi (**dzi-**) sak | bag, sack
kheisi (**dzi-**) *cf Eng* kassie | case
khekhe (**dzi-**) *cf Eng* koek | cake
-khelusa laat afdwaal | coax away
-kheluwa afdwaal, terugval (kerklik) | deviate, desert (one's church)
khemisi (**dzi-**) *cf Eng* apteek | chemist
khemisitiri (*kl* | *cl* **n-**) *cf Eng* chemie, skeikunde | chemistry
khere/kheri (**dzi-**) *cf Afr* kerrie | curry
-kherefa *cf Eng 'care of'* adresseer | put on address
kherefo (**dzi-**) *cf Eng 'care of'* adress | address (of letter)
-khokhekanya (*tr*) ophoop (baie dinge opmekaar) | pile up (many things on top of one another)
khokho (**dzi-**) (*hh*) *cf Eng* kakao | cocoa
khokho (**dzi-**) (*ll*) hoop (klippe, ens) heap (stones, etc)
khokhokho! ideof van klop aan deur | ideoph of knocking at door
khokhonya (*nw* | *n*) (**dzi-**) houtkapper (voël) | woodpecker
-khokhonya (*ww* | *vb*) klop | knock at
kholedzo (**dzi-**) bespotting | mockery
kholidzhi/kholishi (**dzi-**) *cf Eng* kollege | college
kholodiringi (**dzi-**) *cf Eng 'cooldrink'* koeldrank | soft (cold)drink
kholomo (**dzi-**) bees; koei | head of cattle; cow; beast

Figuur 2: Uittreksel uit die *Verbeterde Drietalige Woordeboek* deur Wentzel en Muloiwa (1982: 24)

Sedert die laaste druk in 2009 is hierdie woordeboek egter uit druk. Dit is waarskynlik nie kommersieel haalbaar om dit te herdruk nie, maar dit beteken nie dat die behoefte aan so 'n woordeboek nie meer ter sake is nie. Dit is daarom die moeite werd om ondersoek in te stel na die moontlikheid om met weinig kostes 'n Venda–Afrikaans woordeboek saam te stel en gratis aanlyn aan gebruikers beskikbaar te stel.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

3. Potensiële teikengebruikers

Vir die doel van hierdie bespreking onderskei ons tussen twee tipes gebruikers, te wete direkte en indirekte teikengebruikers. Direkte teikengebruikers is diegene wat die woordeboek gebruik om vertaalekwivalente van byvoorbeeld Vendalemmas in Afrikaans of Engels te vind, of enige variasie van hierdie konfigurasie. So 'n basiese vertaalwoordeboek sal van nut wees vir Afrikaans- of Engelssprekendes wat Venda wil aanleer, of selfs Vendasprekers wat Afrikaans of Engels aanleer. Onderwysstudente sou besonder baat vind by so 'n woordeboek. In die Departement van Hoër Onderwys en Opleiding se hersiene beleid (2015) vir die minimum vereistes vir onderwyskwalifikasies word gestipuleer dat alle afgestudeerde onderwysstudente vaardig moet wees in ten minste een van Suid-Afrika se amptelike tale as 'n taal van onderrig en leer (TvOL). Verder moet elke student kommunikatief vaardig wees in ten minste een ander amptelike Afrikataal. In gevalle waar die taal van onderrig en leer Afrikaans of Engels is, moet die taal van kommunikatiewe vaardigheid 'n Afrikataal wees. Dit is dus moontlik dat 'n student wat Afrikaans as TvOL aanbied, Venda as taal van kommunikatiewe vaardigheid kan kies. 'n Basiese vertaalwoordeboek sal in so 'n geval 'n bruikbare hulpbron wees. Aangesien die woordeboek in hierdie eksperiment nie 'n baie uitgebreide bewerking van lemmas in die vooruitsig stel nie, teiken dit die tipiese 'on the fly'-gebruiker — iemand wat binne 'n bepaalde gebruiksituasie bloot na 'n vertaalekwivalent op soek is. Die beskikbaarheid van so 'n aanlyn hulpbron kan ook die elektroniese voetspoor van 'n minderheidstaal soos Venda vergroot en sodoende bydra tot die status van dié taal as taal van hoër funksies.

Indirekte teikengebruikers stel belang in die databasis wat die woordeboek onderlê, veral as die data in 'n masjienleesbare formaat soos die internasionaal erkende *Extensible Markup Language*, oftewel XML[1], beskikbaar is. In die ontwikkeling van elektroniese hulpbronne, veral vir hulpbronarm tale, is dit belangrik dat hierdie bronne beskikbaar gestel word vir hergebruik. Sulke hulpbronne vorm die basis vir talle Menslike Taaltegnologietoepassings, soos byvoorbeeld masjienvertaling. Die Autshumato-projek[2] is die enigste grootskaalse poging om masjienvertalingsisteme en -hulpbronne vir Suid-Afrikaanse tale daar te stel (cf. McKellar en Groenewald 2012) en bied tans slegs beperkte hulpbronne in die vorm van parallelle korpora en masjienvertaalgeheues vir sekere Suid-Afrikaanse taalpare gratis aan. Afrikaans en Venda as taalpaar word tans nie in hierdie projek gedek nie. Ṋemuṱamvuni (2018) en Moors et al. (2018) bevestig dat die nodige hulpbronne om selfs masjiengesteunde menslike vertaling (oftewel '*machine aided human translation*' soos deur Sager (1994: 326) gedefinieer) met Venda as doeltaal te ontwikkel, nog nie resultate lewer wat met internasionale standaarde vergelyk kan word nie en dat hierdie agterstand grootliks toegeskryf kan word aan 'n tekort aan masjienleesbare hulpbronne.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

4. Beskikbare gratis aanlyn/elektroniese hulpbronne

4.1 Venda: CBOLD (Murphy 1997) — Venda–Engels Woordeboek

In 1994 loods Larry Hyman en John Lowe 'n projek wat ten doel het om 'n aanlyn leksikografiese databasis vir navorsers en leksikograwe daar te stel. As deel van die *Comparative Bantu OnLine Dictionary*-projek (CBOLD) word 'n verskeidenheid hulpbronne in 'n heterogene digitale formaat beskikbaar gestel. 'n Aantal Bantoetaalwoordeboeke word onder 'n oop lisensie aangebied, soos vermeld in die 'Bantuists' Manifesto' wat op die webblad verskyn. Tussen 1994 en 2000 is 'n groot aantal Bantoe-woordeboeke deur CBOLD gedigitiseer en via die projekwebblad vir gebruik en toepassings beskikbaar gestel. Die CBOLD-woordeboeke word in inkonsekwente datastrukture en skemas en in 'n verskeidenheid van formate aangebied (sien ook Eckart et al. 2019: 17.4-17.5). Uiteraard kan hierdie skematiese en tegniese heterogeniteit nie sonder meer vir verdere toepassings gebruik word nie. Transformasie- en kwaliteitsversekeringsmaatreëls is nodig alvorens hierdie waardevolle leksikale databron aktief gebruik kan word.

Een van die beskikbare woordeboeke is 'n Venda-woordeboek (Murphy 1997) wat in die formaat van 'n gewone tekslêer is. Die volgende datavelde word in die woordeboek onderskei: vir naamwoorde word die toonpatroon, die woordklas, die klasnommer waartoe die naamwoord behoort en die Engelse vertaalekwivalent aangedui; vir werkwoorde word die toon van die werkwoordstam aangedui, gevolg deur die woordklas en die vertaalekwivalent in Engels. Hierdie woordeboek vorm die basis van die eksperiment wat in Afdeling 5 beskryf word.

4.2 Gratis Engels–Afrikaanse masjienvertaalsisteme

Masjienvertaling is reeds in die 1940's met behulp van ponskaarte gedoen. Aanvanklik was die kwaliteit van masjienvertaling swak, en selfs tans is sodanige vertaling steeds nie perfek nie, maar soos Groves en Mundt (2015: 113) opmerk, kan die ontwikkeling van kunsmatige intelligensie tot die ontwikkeling van gesofistikeerde vertaalsisteme aanleiding gee. Dit maak daarom sin om die potensiaal van sodanige sisteme te ondersoek.

In hierdie eksperiment word Engels as brug vir die vertaling tussen die brontaal, Venda, en die doeltaal, Afrikaans, gebruik. Gratis beskikbare masjienvertaalsisteme wat vir die vertaling van die Engelse vertaalekwivalente na Afrikaans gebruik sou kon word, is die volgende:

— Google Translate (https://translate.google.com/)

Google Translate is 'n veeltalige neurale masjienvertaaldiens wat deur Google ontwikkel is om teks, dokumente en webwerwe te vertaal. Dit is 'n statisties-

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

gebaseerde sisteem, wat impliseer dat die waarskynlikheid van verskeie korrekte vertaalmoontlikhede bereken word, eerder as wat die sisteem op 'n woord-vir-woord vertaling afgestem is (Groves en Mundt 2015: 113). Tans ondersteun Google Translate 109 tale op verskillende vlakke. Dit bied 'n webwerf-koppelvlak, 'n mobiele toepassing vir Android en iOS en 'n toepassingsprogrammeerkoppelvlak wat ontwikkelaars help om uitbreidings vir webblaaiers ("browsers") en programmatuurtoepassings te bou[3].

— Bing Microsoft® Translator (https://www.bing.com/translator/)

Microsoft® Translator is 'n meertalige wolkvertalingsdiens vir masjienvertalings wat deur Microsoft® gelewer word. Die diens ondersteun tans 87 taalstelsels. Microsoft® Translator bied ook dienste vir teksvertaling via die Translator Text API, wat wissel van 'n gratis vlak wat twee miljoen karakters per maand ondersteun tot betaalde vlakke wat miljarde karakters per maand ondersteun[4].

— Yandex.Translate (https://translate.yandex.com/)

Yandex.Translate, 'n webdiens wat deur die soekenjin en webportaal Yandex verskaf word, is bedoel vir die vertaling van teks of webblaaie. Vertalings is tans beskikbaar in 98 tale. Die stelsel volg 'n hibriede benadering wat statistiese masjienvertaling en neurale masjienvertalingsmodelle kombineer. 'n Woordeboek van enkelwoordvertalings word op grond van die ontleding van miljoene vertaalde tekste gebou. Om die teks te vertaal, vergelyk die algoritme dit eers met 'n databasis van woorde en vergelyk dan die teks met die basistaalmodelle en probeer die betekenis van 'n uitdrukking in die konteks van die teks bepaal[5].

— english-afrikaans.co.za (https://english-afrikaans.co.za/)

Daar is geen verdere beskrywing van hierdie vertaalsisteem beskikbaar nie, en ongelukkig is die vertaalsisteem as sodanig ook aanlyn verwyder sedert die vertalings vir hierdie projek in Julie 2021 afgehandel is.

4.3 Gratis Afrikaanse speltoetser: WSpel

WSpel[6] is tans die omvattendste Afrikaanse speltoetser wat gratis beskikbaar is. Dit voldoen bowendien aan die spelwyse van die jongste *Afrikaanse Woordelys en Spelreëls* (AWS) 2009. Die woordelys bestaan uit ietwat meer as 526,000 woorde en daar is duisende inskrywings wat in Microsoft® Word se AutoCorrect-funksie gebruik kan word. Vir hierdie eksperiment het ons WSpel 15 afgelaai en geïnstalleer.

4.4 Gratis korpussoektogprogramme

Korpussoektogprogramme is onmisbaar vir moderne woordeboeksamestelling,

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

aangesien dit gebruik word om frekwensie-inligting uit 'n korpus te onttrek wat belangrik is vir die saamstel van 'n lemmalys of die aanvulling van 'n bestaande een. In die latere fases van samestelling kan dit onder andere gebruik word vir betekenisonderskeidings tussen polisemiese lemmas en vir die identifisering van gebruiksvoorbeelde en frekwente kollokasies. Vir tale soos Afrikaans, en veral Venda, wat van diakritiese tekens in die ortografie gebruik maak, is dit belangrik dat korpussoektogprogramme hierdie tekens korrek kan hanteer en weergee in die resultate van 'n korpussoektog. Korpussoektogprogramme moet verder die oplaai van eie korpora en die aflaai van soekresultate ondersteun. Vir die doel van hierdie artikel het ons 'n beperkte eksperiment met beide Venda- en Afrikaanse tekste uitgevoer en gevind dat al drie programme hieronder gelys aan bogenoemde vereistes voldoen en dus suksesvol vir korpussoektogte gebruik kan word.

— AntConc: https://www.laurenceanthony.net/software/antconc/
— LancsBox: http://corpora.lancs.ac.uk/lancsbox/
— Voyant Tools: https://voyant-tools.org/

4.5 Venda-korpus

Die gebruik van korpora is reeds standaardpraktyk in moderne woordeboeksamestelling. Korpusdata word gebruik om lemmalyste op te stel, frekwensie-inligting te bekom en is 'n potensiële bron van gebruiksvoorbeelde. 'n Tipiese korpussoektog na sleutelwoorde in konteks ('n sogenaamde KWIC oftewel "Keyword in Context"-soektog, Afrikaans SWIK-soektog) met konkordansielyne as resultaat, is 'n onmisbare bron vir die identifisering van alle moontlike betekenisse van 'n bepaalde lemma. Vir die doel van hierdie eksperiment is 'n Venda-korpus van 1.4 miljoen ortografiese woorde ("tokens") gebruik. Hierdie korpus vorm deel van 'n projek van die *South African Centre for Digital Language Resources* (SADiLaR), wat die daarstel van elektroniese hulpbronne, spesifiek vir die Afrikatale, ten doel het. Dit is 'n rou korpus, sonder enige annotasie van byvoorbeeld woordkategorieë, en bestaan uit gedigitiseerde tekste oor 'n wye verskeidenheid genres, maar sluit nie 'n gesproke komponent in nie. Dit is heel waarskynlik nie gebalanseerd nie, maar om Atkins et al. (1992: 14) aan te haal:

> In 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. It would be shortsighted indeed to wait until one can scientifically balance a corpus before starting to use one, and hasty to dismiss the results of corpus analysis as 'unreliable' or 'irrelevant' simply because the corpus used cannot be proved to be 'balanced'.

Hierdie opmerking is veral geldig ten opsigte van die Afrikatale, spesifiek vir data-arm tale soos Venda.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

4.6 Gratis program vir woordeboeksamestelling: *Lexonomy*

Die navorsingsprojek genaamd *European Lexicographic Infrastructure* (ELEXIS) het ten doel om die tale van Europa te verbind deur die woordeboeke wat vir hierdie tale bestaan te standaardiseer, digitiseer en met mekaar te vergelyk (ELEXIS 2020: 1):

> More people than ever before use dictonaries. Not so much printed ones, but dictonary data are now built into mobile phones, software, systems and are in use all the time. New and updated dictonaries are badly needed fast (ELEXIS 2020: 1).

Om daardie doel te bereik, het die projek 'n verskeidenheid hulpbronne en gereedskapstelle ontwikkel en gratis vir leksikograwe, ontwikkelaars van Mensliketaaltegnologie (MTT) en taalgebruikers beskikbaar gestel. Een van die mees prominente hulpmiddels is *Lexonomy* — 'n wolkgebaseerde, oopbron woordeboekskrywer en -publiseerder[7].

Die ontwikkelaars van *Lexonomy* wou 'n hulpbron skep wat geen installasie, programmeringskennis of duur kostes dra om 'n basiese aanlyn woordeboek te kan skep nie en só meer gebruikers in staat stel om hul eie woordeboeke op te stel (Měchura 2017: 662). Die platform kan gebruik word om 'n een- of meertalige woordeboek op te stel, te formateer, aan te pas en uiteindelik aanlyn te publiseer. Dit is verder ook moontlik om 'n kollektiewe projek op te stel sodat 'n groep mense saam aan 'n woordeboek kan werk. Die woordeboek bly 'n privaat projek totdat die eienaar dit publiek maak. *Lexonomy* se webblad word dan 'n platform waarop die woordeboek gestoor en gebruik word en van waar dit gedeel word sodat enige gebruiker toegang kan kry.

Alhoewel *Lexonomy* gebruik kan word om 'n nuwe woordeboek van die begin af te ontwikkel en leksikograwe toerus om elke leksikale item stelselmatig van tradisionele datatipes soos die hoofwoord of lemma, woordsoort, definisie en gebruiksvoorbeeld te voorsien, is dit ook moontlik om bestaande data (semi-)outomaties in die sisteem in te voer (Měchura 2017: 664). Die data kan enigiets van 'n eenvoudige woordelys tot 'n uitgebreide veeltalige woordeboek wees, solank dit aan basiese vereistes vir XML formaat voldoen en elke datatipe met die toepaslike merkers geïdentifiseer kan word. As 'n bestaande datastel opgelaai word, kan dit steeds in naredigering deur 'n leksikograaf aangepas/ uitgebrei word of volgens 'n nuwe protokol geformateer word. Op dieselfde manier kan 'n woordeboek wat in *Lexonomy* opgestel is, ook in XML formaat afgelaai word sodat dit op 'n ander platform versprei of verder gebruik kan word.

Hierdie twee funksies (die op- en aflaai na en van die XML-databasis) is veral belangrik omdat dit die volhoubare ontwikkeling en gebruik van die resulterende woordeboek verseker. Die XML-formaat kan geredelik deur ander MTT-toepassings hergebruik word omdat die struktuur en merkers vooraf bepaal en in 'n stylgids beskryf word. Elke leksikale item in die woordeboek

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

word volgens dieselfde voorafbepaalde protokol behandel en spesifieke datavelde vir elke item kan dus onttrek word, afhangend van die toepassing waarvoor die data aangewend sal word, byvoorbeeld as afrigtingsdata vir masjienvertaling en inligtingonttrekking, as die basis van 'n woordnet of 'n domeinspesifieke woordeboek. Die XML-struktuur beteken ook dat 'n ontwikkelaar wat nie noodwendig 'n spreker van die taal is nie die datavelde in elke leksikale item kan identifiseer en verder manipuleer vir ander toepassings. So bevorder aanlyn leksikografie ook die ontwikkeling van tale met minder hulpbronne in die MTT-arena.

Lexonomy is al in verskeie soortgelyke projekte aangewend. Stemle et al. (2019: 537-546) beskryf die skep van 'n neologismewoordeboek vir 'n standaardvariant van Duits en hoe *Lexonomy* toegepas word om die woordeboek maklik leesbaar en, belangriker nog, maklik aanpasbaar te maak sodat lemmas gereeld bygevoeg kan word. In hierdie projek word bestaande korpora en woordeboeke ook ingespan om die nuwe woordeboek volgens die sogenaamde "eenkliek-woordeboek paradigma" (oftewel die "one-click dictionary paradigm") saam te stel. Daarin lê natuurlik die grootste verskil tussen die eksperiment wat hier beskryf word en hierdie internasionale projek — vir Venda bestaan daar nog weinig hulpbronne om in die ontwikkelingspyplyn te gebruik.

Bartolomé-Díaz en Frontini (2020: 62-68) gebruik dieselfde komponente om 'n Frans–Spaans tweetalige woordeboek saam te stel, maar fokus hul studie op die kennis en vaardighede wat die samesteller nodig het om so 'n taak suksesvol uit te voer. Die gevolgtrekking is dat *Lexonomy* verder verbeter kan word om byvoorbeeld meer metadata saam met die woordeboek te versamel en om hiperskakels by elke inskrywing toe te laat. Hulle sluit ook by die werk van Jakubíček et al. (2018: 65-68) aan wat op uitdagings in die formatering van woordeboeke volgens streng internasionale standaarde en tydens naredigering wys, maar albei is dit eens dat hierdie uitdagings maklik deur 'n bekwame leksikograaf en versigtige outomatisering van dele van die formateringstappe in *Lexonomy* omseil kan word.

5. Eksperiment

Hierdie afdeling beskryf die stappe wat gevolg is om die effektiwiteit van woordeboeksamestelling met die beskikbare hulpbronne, soos reeds bespreek, te toets. Figuur 6 aan die einde van die afdeling verteenwoordig 'n grafiese voorstelling van die prosedure wat in hierdie eksperiment gevolg is.

5.1 Voorafredigering

Die Engels–Venda woordeboek is eers afgelaai en na 'n eenvoudige sigblad in Microsoft® Excel oorgedra. Deur die data in kolomme te verdeel en 'n unieke ID aan elkeen van die inskrywings toe te ken, kon die navorsingspan maklik

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

besluit watter velde in watter stappe gebruik sou word. Die Engelse vertalings kon byvoorbeeld maklik van die Venda-ekwivalente geskei word om in die masjienvertalingsproses te gebruik.

Vir die doeleindes van hierdie eksperiment is 10% van die woordeboekinskrywings ewekansig uitgesoek. Geslote woordklasse is egter uitgesluit, dus is 800 ewekansig gekose terme uit die volgende 4 woordsoortklasse ingesluit: naamwoorde, werkwoorde, bywoorde en adjektiewe. Hierdie beginsel is uit masjienleer geleen. Die verdeling van 'n korpus in 3 dele waar 80% vir die afrigting van 'n model gebruik word, 10% vir die evaluering van die model en 10% vir 'n ontwikkelingstel word algemeen aanvaar. Hierdie verdeling word ook in Jurafsky en Martin (2009: 187) aanbeveel wanneer hulle oor evaluering en foutanalise skryf:

> We train our tagger on the training set. Then we use the development test set (also called a dev-test set) to perhaps tune some parameters, and in general decide what the best model is. Once we come up with what we think is the best model, we run it on the (hitherto unseen) test set to see its performance. We might use 80% of our data for training and save 10% each for dev-test and test. Why do we need a development test set distinct from the final test set? Because if we used the final test set to compute performance for all our experiments during our development phase, we would be tuning the various changes and parameters to this set.

5.2 Outomatiese vertaling van Engels na Afrikaans

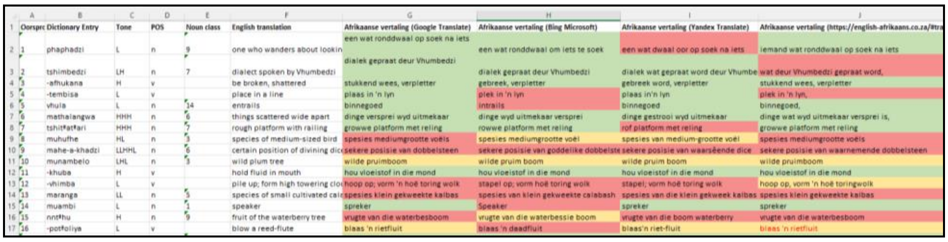
Tydens die voorafredigering van die Engelse weergawe vir outomatiese vertaling van Engels na Afrikaans is die volgende afkortings outomaties met 'n vind-en-vervang-proses na hul volle vorm verander:

s/t na *something*
s/o na *someone*
w/ na *with*
everything/body na *everything/everybody*

Die Engelse data uit die ontwikkelingstel (oftewel "training set", soos bo beskryf), is hierna outomaties na Afrikaans vertaal met die vier masjienvertaalsisteme wat in Afdeling 4.2 beskryf is. Die resulterende Afrikaanse vertalings is daarna ook in die sigblad gevoeg om nou 'n ontwikkelingstel met een Engelse, een Venda en vier (moontlike) Afrikaanse vertalings in te sluit.

Die volgende stap was om die vertalings te evalueer om sodoende 'n enkele geskikte Afrikaanse vertaling te kies. Hiervoor is menslike intervensie nodig. Vir die handmatige evaluering is punte vir die vertalings deur elke vertaalsisteem toegeken. 'n Perfekte vertaling het twee punte gekry terwyl aan 'n vertaling met bv. 'n spelfout, foutiewe woordorde, foutiewe ortografie, ens. slegs een punt toegeken is. Onbruikbare vertalings het geen punt ontvang nie.

Drie eerstetaalsprekers van Afrikaans het die 800 terme van die ontwikkelingstel geëvalueer en verskille in puntetoekenning met mekaar bespreek om op 'n finale punt te besluit. Afdeling 6 beskryf die evaluering van die masjienvertaalsisteme en die foutanalise breedvoerig en Figuur 3 gee 'n blik op die sigblad waarin evaluering gedoen is. Voorwaardelike formatering ("conditional formatting"), 'n funksie in Microsoft® Excel, word gebruik om aan te dui watter punt elke vertaling ontvang het — groen dui op 'n korrekte vertaling, geel op 'n vertaling wat met 1 gemerk is en dus nog menslike insette sal nodig hê om aanvaarbaar te wees, en rooi dui op 'n onbruikbare vertaling. Deur hierdie kleure aan te bring, kon die drie evalueerders maklik sien waar hul puntetoekenning verskil ten einde dit te bespreek en op te los.



	A	B	C	D	E	F	G	H	I	J
1	Oorspr	Dictionary Entry	Tone	POS	Noun class	English translation	Afrikaanse vertaling (Google Translate)	Afrikaanse vertaling (Bing Microsoft)	Afrikaanse vertaling (Yandex Translate)	Afrikaanse vertaling (https://english-afrikaans.co.za/#tra
2	1	phaphadzi	L	n	9	one who wanders about lookin	een wat ronddwaal op soek na iets	een wat ronddwaal om iets te soek	een wat dwaal oor op soek na iets	iemand wat ronddwaal op soek na iets
3	2	tshimbedzi	LH	n	7	dialect spoken by Vhumbedzi	dialek gepraat deur Vhumbedzi	dialek gepraat deur Vhumbedzi	dialek wat gepraat word deur Vhumbe	wat deur Vhumbedzi gepraat word,
4	3	-afhukana	H	v		be broken, shattered	stukkend wees, verpletter	gebreek, verpletter	gebreek word, verpletter	stukkend wees, verpletter
5	4	-tembisa	L	v		place in a line	plaas in 'n lyn	plek in 'n lyn	plaas in'n lyn	plek in 'n lyn,
6	5	vhula	L	n	14	entrails	binnegoed	intrails	binnegoed	binnegoed,
7	6	mathalangwa	HHH	n	6	things scattered wide apart	dinge versprei wyd uitmekaar	dinge wyd uitmekaar versprei	dinge gestrooi wyd uitmekaar	dinge wat wyd uitmekaar versprei is,
8	7	tshit*at*ari	HHH	n	7	rough platform with railing	growwe platform met reling	rowwe platform met reling	rof platform met reling	growwe platform met reling
9	8	muhufhe	HL	n	3	species of medium-sized bird	spesies mediumgrootte voëls	spesies mediumgrootte voël	spesies van medium-grootte voël	spesies mediumgrootte voëls
10	9	mahe-a-khadzi	LLHHL	n	6	certain position of divining dic	sekere posisie van dobbelsteen	sekere posisie van goddelike dobbelste	sekere posisie van waarsêende dice	sekere posisie van waarnemende dobbelsteen
11	10	munambelo	LHL	n	3	wild plum tree	wilde pruimboom	wilde pruim boom	wilde pruim boom	wilde pruimboom
12	11	-khuba	H	v		hold fluid in mouth	hou vloeistof in die mond	hou vloeistof in die mond	hou vloeistof in die mond	hou vloeistof in die mond
13	12	-vhimba	L	v		pile up; form high towering clo	hoop op; vorm 'n hoë toring wolk	stapel op; vorm hoë toring wolk	stapel; vorm hoë toring wolk	hoop op, vorm 'n hoë toringwolk
14	13	maranga	LL	n	5	species of small cultivated cal	spesies klein gekweekte kalbas	spesies van klein gekweekte calabash	spesies van die klein gekweek kalbas	spesies klein gekweekte kalbas
15	14	muambi	L	n	1	speaker	spreker	Speaker	spreker	spreker
16	15	nnt*hu	H	n	9	fruit of the waterberry tree	vrugte van die waterbesboom	vrugte van die waterbessie boom	vrugte van die boom waterberry	vrugte van die waterbesboom
17	16	-pot*oliya	L	v		blow a reed-flute	blaas 'n rietfluit	blaas 'n daadfluit	blaas'n riet-fluit	blaas 'n rietfluit

Figuur 3: Basisdokument vir evaluering

5.3 Verbetering van vertalings met WSpel

Een van die doelstellings van hierdie eksperiment was om die samestellingsproses sover moontlik te automatiseer en die beskikbare hulpbronne optimaal te benut. Die navorsingspan het dus besluit om 'n speltoetser as 'n eerste stap in die redigering van die Afrikaanse vertalings in te span. Afdeling 4.3 gee besonderhede oor WSpel 15, 'n gratis Afrikaanse hulpbron. WSpel se hulpgids beskryf omvattende stappe vir installasie en gebruik met 'n paar weergawes van Microsoft® Office, maar die installasieproses is omslagtig en die beskrywing is met tye moeilik om te volg. Dit is ook net in Afrikaans beskikbaar en dus nie geskik vir leksikograwe wat nie Afrikaans magtig is nie.

Die navorsingspan het die Afrikaanse vertalings wat in die vorige stap met 'n 1 en 2 gemerk is (wat dus óf net so in die nuwe woordeboek ingesluit kan word, óf net klein veranderinge nodig het — sien Afdeling 5.2) met WSpel getoets om sodoende vas te stel of 'n speltoetser die taak van 'n redigeerder kan vergemaklik. In hierdie eksperiment het WSpel egter geen spelfoute gevind nie en kon die span dus nie die effektiwiteit daarvan as deel van die samestellingsproses behoorlik toets nie.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

5.4 *Lexonomy*-opstelling en keuse van datavelde

Lexonomy (sien Afdeling 4.6 vir 'n volledige beskrywing) is bedoel om aanpasbare oplossings vir verskillende tipes projekte vir woordeboeksamestelling te bied. Dit is dus moontlik om vooraf 'n stylblad met verpligte en opsionele velde, verskillende formaterings en teksgroottes op te stel sodat enige nuwe inskrywings daarvolgens ontwikkel word. Dit is veral nodig wanneer 'n span leksikograwe saam aan die ontwikkeling van 'n woordeboek werk en waar inskrywings een vir een bygevoeg word.

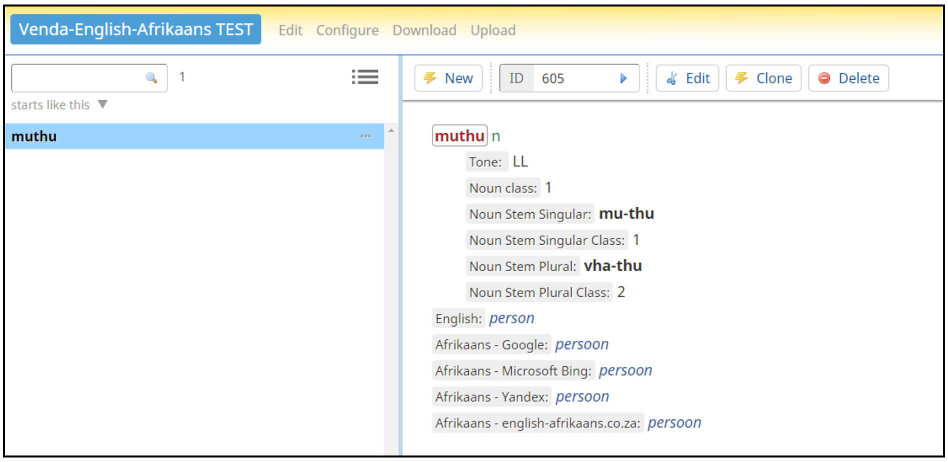
Wat hierdie sagteware egter vir ons projek uiters geskik maak, is die feit dat 'n bestaande datastel in XML-formaat opgelaai kan word en in *Lexonomy* bygewerk, verander of vir gebruik beskikbaar gestel kan word. Aangesien ons juis poog om menslike intervensie tot 'n minimum te beperk, en omdat ons reeds 'n groot deel van die voorafredigering, vertaling en verbeterings in Microsoft® Excel gedoen het, wou ons hierdie eienskap van *Lexonomy* maksimaal benut om 'n bruikbare e-woordeboek saam te stel wat later maklik bygewerk en verbeter kon word.

Die navorsingspan het besluit om vir die doeleindes van hierdie eksperiment te hou by die datavelde en struktuur wat in die oorspronklike woordeboek gebruik is en ook die opstelling in *Lexonomy* so te doen. Die relevante data is vervolgens uit Microsoft® Excel na 'n tekslêer oorgedra en die nodige XML-merkers is by elke kolom gevoeg. Die volgende datavelde is dus ingesluit:

— unieke ID wat aan die begin van die proses aan elkeen van die 800 terme as identifiseerder toegeken is;
— hoofwoord/lemma in Venda tesame met die klasnommer;
— Engelse vertaling; en
— Afrikaanse vertaling.

Omdat *Lexonomy* so maklik aanpasbaar is, kan addisionele datavelde soos definisies, sinonieme, gebruiksnotas, en so meer, baie maklik op 'n later stadium bygevoeg word.

Die 800 terme wat in hierdie eksperiment gebruik is, is vervolgens as 'n eerste toets van dié konsep in *Lexonomy* opgestel. Figuur 4 wys 'n skermskoot van 'n inskrywing in *Lexonomy* met al die addisionele inligting uit die Venda-woordeboek. Die span het egter besluit om vir hierdie eksperiment net die inligting wat ook in die woordeboek van Murphy (1997) opgeneem is, te vertoon om direkte vergelykings tussen die oorspronklike gedrukte weergawe en die nuwe elektroniese weergawe wat hier geskep is, te kan tref. Hierdie uiteindelike uitleg word in Figuur 5 gewys.



muthu

muthu n

Tone: LL

Noun class: 1

Noun Stem Singular: **mu-thu**

Noun Stem Singular Class: 1

Noun Stem Plural: **vha-thu**

Noun Stem Plural Class: 2

English: *person*

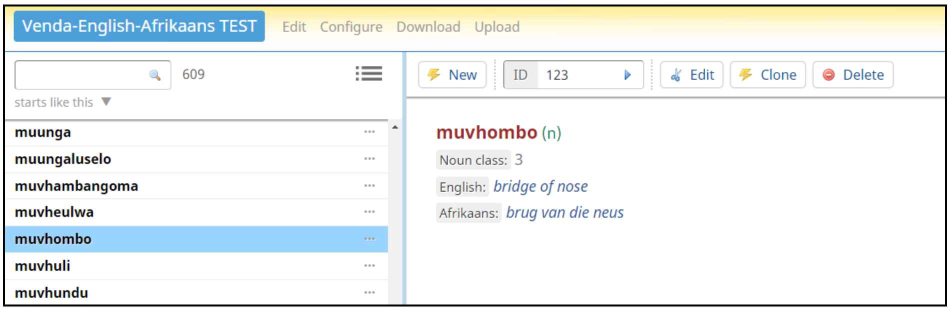
Afrikaans - Google: *persoon*

Afrikaans - Microsoft Bing: *persoon*

Afrikaans - Yandex: *persoon*

Afrikaans - english-afrikaans.co.za: *persoon*

Figuur 4: Voorbeeld van 'n prototipiese inskrywing



muunga
muungaluselo
muvhambangoma
muvheulwa
muvhombo
muvhuli
muvhundu

muvhombo (n)

Noun class: 3

English: *bridge of nose*

Afrikaans: *brug van die neus*

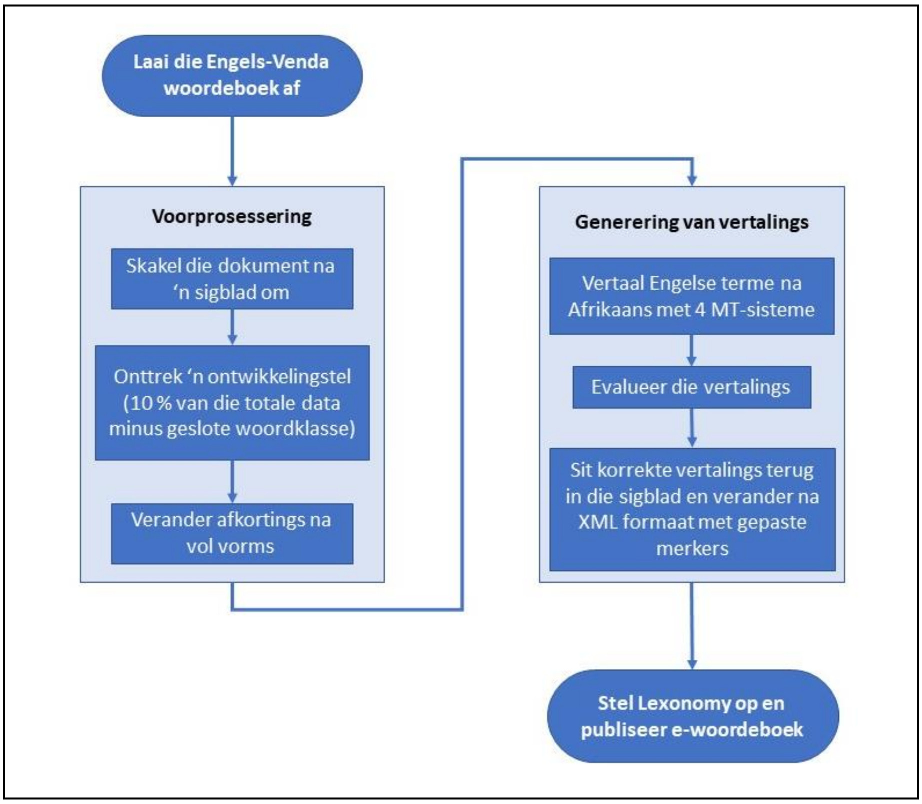
Figuur 5: Die finale uitleg van 'n inskrywing

5.5 Uitbreiding van die lemmalys

In Afdeling 3 hierbo is aangedui dat een groep teikengebruikers aanleerders van Venda as tweede of derde taal is. Dit is daarom belangrik dat lemmas met 'n hoë gebruiksfrekwensie in die lemmalys opgeneem word. As deel van die eksperiment is die Venda-korpus wat in Afdeling 4.5 hierbo beskryf is, gebruik om 'n frekwensielys op te stel. *LancsBox* is as korpusnavraagprogrammatuur gebruik. Van die 500 mees frekwente woorde wat deur *LancsBox* se frekwensiesoektog opgelewer is, kom slegs 288, d.w.s. slegs 57.6% in Murphy (1997) se woordeboek voor. Hierdie 500 woorde maak 12% van die totale korpus uit en dit is daarom belangrik dat hulle as lemmas in die woordeboek opgeneem word. Van die 27 mees frekwente woorde, wat tipies grammatiese formatiewe soos kongruensiemorfeme, ander werkwoordprefikse en naamwoordelike prefikse insluit, is nie een in die woordeboek opgeneem nie. Van die mees frekwente woorde met leksikale betekenis wat nie in die woordeboek verskyn nie, is die volgende: *ṅwana* 'child' (frekwensierangorde 82), *-ḓivha* 'weet, ken' (frekwensierangorde 94), *buthano* 'byeenkoms' (frekwensierangorde 290) en *-bvela* 'uitkom, kom na' (frekwensierangorde 446). Deur die oorblywende 212 mees frekwente lemmas by die lemmalys te voeg, verhoog dit die kans op 'n suksesvolle soektog, aangesien hoë-frekwensie woorde juis dié is wat deur aanleerders nageslaan word.

Tydens die eksperiment het dit duidelik geword dat sodanige aanvulling van die lemmalys die intervensie van die leksikograaf nodig het. Daar bestaan in die eerste plek 'n hoë graad van homografie ten opsigte van die grammatiese formatiewe en die verskillende betekenisse/betekenisfunksies sal deur die leksikograaf ontrafel moet word. Tweedens sal die betekenisparafrase van die hoë-frekwensie woorde wat tot die lemmalys bygevoeg word ook deur die leksikograaf voorsien moet word.

Figuur 6 gee 'n grafiese voorstelling van die stappe in hierdie eksperiment.


Laai die Engels-Venda woordeboek af
Voorprosessering
Skakel die dokument na 'n sigblad om
Onttrek 'n ontwikkelingstel (10 % van die totale data minus geslote woordklasse)
Verander afkortings na vol vorms
Generering van vertalings
Vertaal Engelse terme na Afrikaans met 4 MT-sisteme
Evalueer die vertalings
Sit korrekte vertalings terug in die sigblad en verander na XML formaat met gepaste merkers
Stel Lexonomy op en publiseer e-woordeboek

Figuur 6: 'n Diagrammatiese voorstelling van die eksperiment

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

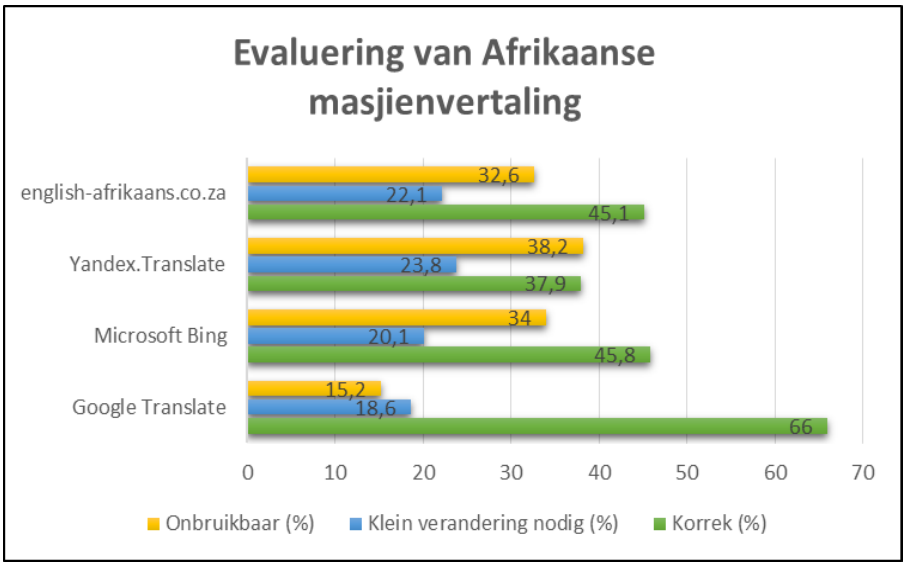
6. Evaluering

6.1 Masjienvertaling

Uit die handmatige evaluering van die masjienvertaling van Engels na Afrikaans blyk dit duidelik dat die vier sisteme nie almal ewe effektief is nie. Die tabel (Tabel 1) onder gee enkele algemene indrukke wat in Afdeling 6.2 met voorbeelde aangevul word.

	Google Translate	Microsoft® Bing	Yandex. Translate	english-afrikaans.co.za
Korrekte vertaling	528	366	303	361
Klein verandering nodig	149	161	190	177
Onbruikbare vertaling	122	272	306	261
Onseker	1	1	1	1
Totaal	**800**	**800**	**800**	**800**

Tabel 1: Samevatting van die handmatige evaluering



Figuur 7: Grafiese voorstelling van die samevatting van die handmatige evaluering

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

Google Translate lewer dus die meeste vertalings wat direk in 'n woordeboek ingesluit kan word sonder dat enige menslike intervensie nodig is, met 528 (66 %) van die inskrywings wat as korrek gemerk is. *Yandex.Translate* blyk die meeste naredigering te verg omdat 190 (23.8 %) van die inskrywings klein veranderinge nodig het en 306 (38.2 %) heeltemal onbruikbaar is en dus deur 'n linguis voorsien sal moet word.

Dit is ook interessant dat 613 van die 800 gevalle (77 %) ten minste een korrekte vertaling bevat en dat al vier vertalers in 168 gevalle (21 %) 'n korrekte (maar nie noodwendig dieselfde) vertaling lewer en dus 'n hoë sekerheidsgraad behaal. Die voordeel van 'n elektroniese woordeboek is natuurlik dat die samestellers nie hier 'n keuse vir net een vertaling hoef te maak nie en selfs al vier kan insluit omdat die formaat nie soveel beperkings op die artikellengte stel nie.

In 26 gevalle is ten minste een van die vertalings met 'n 1 gemerk, wat beteken dat net klein veranderings deur 'n leksikograaf of taalkenner nodig sal wees om 'n aanvaarbare vertaling te lewer wat ook in die woordeboek gevoeg kan word.

Die volgende afdeling bespreek van die vertaaluitdagings in meer besonderhede.

6.2 Vertaaluitdagings

6.2.1 (Semi-)outomatiese korreksies

Sekere uitdagings kan (semi-)outomaties reggestel word, bv. woordorde, ortografie (koppeltekens, hoofletters, spasiëring) en vertalings van meervoud versus enkelvoud.

Woordorde

Die woordorde in baie van die Afrikaanse vertalings volg die woordorde van die Engelse weergawe, bv. *become wealthy* word in vertaalsisteem D as *word ryk* vertaal, terwyl vertaalsisteme A en B die woordvolgorde korrek as *ryk word* vertaal. Nog 'n voorbeeld is: *bring nearer* wat deur A en B korrek as *nader bring* vertaal word, terwyl C en D die volgorde omruil *bring nader.* Hierdie woordvolgordeprobleem duik slegs by werkwoorde op, in gevalle waar die Engelse werkwoord uit meer as een woord bestaan. By nadere beskouing het dit duidelik geword dat in die meerderheid van die ongeveer 100 gevalle waar geen van die Afrikaanse vertalings suksesvol was nie, woordorde die probleem is (30 gevalle). Vir Afrikaanstalige gebruikers van die woordeboek behoort dit geen probleem te wees nie.

Ortografie

Daar is heelwat probleme rakende ortografiese aangeleenthede soos bv.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

die gebruik van koppeltekens al dan nie, en ook die vas- en losskryf van woorde. Enkele voorbeelde is: *akasia-boom* vs. *akasiaboom*, *Venda-taal* vs. *Vendataal* en *swart mamba slang* vs. *swart mamba-slang.* Dié ortografiese probleem kom veral in die geval van eiename soos boom- en plantname voor, maar kan maklik met behulp van 'n Afrikaanse speltoetser opgelos word. Daar is bevind dat slegs 16 van die ongeveer 100 gevalle waar geen van die Afrikaanse vertalings suksesvol was nie, met koppeltekens en die vas- en losskryf van woorde te doen het.

In sommige gevalle kom 'n vertaalsisteem met 'n hoofletter vorendag, bv. *gaste* vs. *Gaste.* 'n Semi-outomatiese vind-en-vervang proses is 'n maklike oplossing, maar daar moet versigtig te werk gegaan word sodat eiename nie outomaties na kleinletters verander word nie.

Spasiëring is deurgaans 'n probleem in vertaalsisteem C wanneer dit kom by 'n spasie tussen die lidwoord *'n* en die voorafgaande woord, bv. *braai oor 'n oop vuur, opgesluit in 'n klein hok.* In dié gevalle is 'n outomatiese vind-en-vervang prosedure die aangewese oplossing.

Vertalings van meervoud vs. enkelvoud

Die Engelse woord *species* kom slegs in die meervoud voor, maar die korrekte weergawe in Afrikaans is die enkelvoud *spesie,* bv. die korrekte vertaling van *species of medium-sized bird* sou wees *spesie mediumgrootte voël.* Die vier vertaalsisteme is nie konsekwent met die enkelvoud- en meervoud-vertalings nie. 'n Semi-outomatiese vind-en-vervang prosedure is 'n maklike oplossing van die probleem. Dieselfde geld vir 'n voorbeeld soos *fruit* in *fruit of muhukhuma.* Volgens die *Macmillan English Dictionary for Advanced Learners* (2002: 571) is die meervoud van *fruit* óf *fruit* óf *fruits* en dit is waarskynlik die rede waarom die vertaalsisteme inkonsekwent is met die Afrikaanse vertaling. Na aanleiding van die Venda enkelvoud-naamwoordklas 9 van *muhukhuma* sou die korrekte vertaling dus *vrug* en nie *vrugte* wees nie. Weereens is 'n semi-outomatiese vind-en-vervang prosedure is 'n maklike oplossing. Dit is slegs van toepassing op 11 van die ongeveer 100 gevalle waar geen van die Afrikaanse vertalings suksesvol was nie.

Hoflikheidsvorm "u"

In sommige Afrikaanse vertalings word die hoflikheidsvorm "u" gebruik, bv. in *inciting others to do something that causes discord, while staying uninvolved oneself* wat deur sisteme A en D vertaal word as *ander aan te spoor om iets te doen wat onenigheid veroorsaak, terwyl u onbetrokke bly.* Sulke gevalle kan ook met 'n semi-outomatiese vind-en-vervang prosedure benader word.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

6.2.2 Menslike intervensie

Menslike intervensie is nodig in gevalle waar konteks (polisemie), kulturele begrippe en domeinspesifieke terme 'n uitdaging is. Korreksies moet in sulke gevalle handmatig aangebring word.

Konteks

Daar is verskeie gevalle van polisemie opgemerk waar 'n woord verskeie betekenisse in verskillende kontekste het, soos in die voorbeeld:

> *bar or pole for barring cattle kraal gate > kroeg of paal vir die versperring van beeskraalhek > kroeg of paal vir die versperring van beeste kraal hek > bar of paal behalwe vir beeste kraal hek bar of paal vir blokkeer beeskraal hek.*

In 'n ander konteks sou *bar* met *kroeg* vertaal kon word, maar in hierdie spesifieke konteks is die betekenis *paal* die aangewese een in die konteks van *beeskraalhek.*

Dieselfde geld vir die polisemiese Engelse werkwoorde *pull, draw* wat in sisteme A en B as *trek, teken* vertaal word. In die gegewe konteks is slegs *trek* korrek. 'n Soortgelyke voorbeeld is *rim* en *edge* wat albei as *rand* in Afrikaans vertaal word. Die Engelse weergawe *make a rim or edge on a basket* word deur al die sisteme verkeerdelik as *maak 'n rand of rand op 'n mandjie* vertaal.

Kultuurgebonde begrippe

Een van die Engelse definisies van die Venda-werkwoordstam *-xa* word gegee as *lose all counters in mufuvha game*. Agtergrondkennis (soos gevind by https://www.bead.game/games/traditional/mefuvha) is nodig om die korrekte vertaling in Afrikaans te identifiseer, naamlik *tellers* en nie *toonbanke* nie. Vertaalsisteme B en C gee wel die korrekte vertaling.

Die naamwoord *chief* in die volgende voorbeeld is ook kultuurgebonde: *hut of chief's uncle, brother, son.* Die korrekte vertaling is dus *hut van hoofman se oom, broer, seun.* Sisteme B en D gee die korrekte vertaling as *hoofman*, terwyl sisteme A en C onderskeidelik *owerste* en *hoof* as vertaling gee wat wel in 'n ander konteks korrek sou wees.

Hoewel kultuurgebonde begrippe hoogswaarskynlik 'n baie lae frekwensie in Venda-tekste het, het hulle tog kultuurhistoriese waarde.

Domeinspesifieke terme

Leemtes in die leksikons van die vertaalsisteme kom na vore in die geval van domeinspesifieke terme, soos bv. plant- en voëlname, waar dikwels 'n letterlike vertaling gedoen word. Die term *Cape robin* word deur drie van die vier vertaalsisteme as *Kaapse robin* vertaal terwyl een sisteem slegs die Engelse term weergee. Die korrekte Afrikaanse term is *janfrederik.* 'n Voorbeeld van 'n plantnaam is *cabbage tree* wat deur al vier sisteme letterlik vertaal word as *koolboom* of *kool boom* in plaas van *kiepersol*.

Meerdere vertaalekwivalente

Dit is interessant om op te merk dat volgens 'n woordeboek soos Pharos se *Verklarende Afrikaanse woordeboek* woorde soos *os* en *bees* as ekwivalent beskou word, en ook dat *blom* as ekwivalent vir *blossom* gegee word. Dit is derhalwe nie vreemd dat die vertaalsisteme daarmee akkoord gaan nie.

Afwesigheid van korrekte vertaalekwivalente

Indien daar geen korrekte vertaalekwivalent bestaan nie, sal die leksikograaf uiteraard 'n vertaalekwivalent moet verskaf, soos in die geval van die werkwoord *khakhamedza* met die Engelse vertaling *take aback*. Al vier Afrikaanse vertalings wat verskaf word, is foutief, naamlik *skrik, neem terug* en *neem uit die veld geslaan* (laasgenoemde word deur twee van die vertaalsisteme verskaf). Die korrekte vertaling sou wees *verstom* of *uit die veld slaan*.

7. Gevolgtrekking

7.1 Gevolgtrekkings en samevatting

Die bevindinge van ons aanvanklike ondersoek na die beskikbaarheid van meertalige woordeboeke vir Afrikatale met skaars hulpbronne, soos vir Venda, het daarop gedui dat die enigste Venda–Engels–Afrikaans woordeboek al vir 'n geruime tyd uit druk is. Die behoefte aan so 'n woordeboek het intussen ontstaan as gevolg van twee tipes gebruikers, naamlik direkte en indirekte teikengebruikers. Direkte teikengebruikers sluit taalaanleerders soos onderwysstudente in, terwyl indirekte teikengebruikers daarna streef om die woordeboekdata te gebruik om die taal tegnologies te ontwikkel, onder andere vir doeleindes van masjienvertaling.

In hierdie artikel word die optimale benutting van gratis elektroniese/aanlyn hulpbronne vir die saamstel van 'n bruikbare drietalige e-woordeboek vir Venda, Engels en Afrikaans wat mettertyd maklik bygewerk kan word, ondersoek. Die benadering wat gevolg is, behels 'n eksperiment waarin die samestellingsproses so ver moontlik geoutomatiseer is om besparing in terme van tyd en mens-ure teweeg te bring. Engels word as 'n brug vir die vertaling tussen die brontaal, Venda, en die doeltaal, Afrikaans, gebruik. Die gratis beskikbare hulpbronne wat gebruik is, sluit in 'n Venda–Engels woordeboek, vier Engels–Afrikaans masjienvertaalsisteme, 'n Afrikaanse speltoetser, korpusondersoekprogramme en 'n program vir woordeboeksamesteling. Hierdie eksperiment is op 10% ewekansig uitgesoekte woordeboekinskrywings gebaseer wat vier woordsoortklasse insluit, naamlik naamwoorde, werkwoorde,

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

bywoorde en adjektiewe. Geslote woordklasse is uitgesluit. Die handmatige evaluering deur eerstetaalsprekers is ook op hierdie data uitgevoer.

Die algemene bevindinge van die eksperiment is dat daar — soos te wagte — sekere beperkings op so 'n semi-outomatiese proses met gratis hulpbronne is, wat 'n sekere mate van menslike intervensie verg. Hoewel die saamgestelde e-woordeboek nie as 'n finale produk beskou kan word nie, bied die wolkgebaseerde, oopbron woordeboekskrywer en -publiseerder *Lexonomy* die geleentheid vir menslike insette soos deur byvoorbeeld leksikograwe, linguiste, ens. om die nodige bywerkings op 'n gebruikersvriendelike wyse te doen. Dit is te danke aan die aanpasbaarheid en maklike uitleg van *Lexonomy*. Verdere woordklasse, veral geslote woordklasse, kan met behulp van 'n grammatika soos Poulos (1990) se *A Linguistic Analysis of Venda* vergelyk word, en deur menslike intervensie aangevul word, bv. die verskillende tipes voornaamwoorde. Dit is interessant om op te merk dat 'n geslote woordklas soos voegwoorde goed verteenwoordig is in die Venda–Engels Woordeboek (Murphy 1997), naamlik 27 voegwoorde altesaam, terwyl Poulos (1990) slegs die 15 mees frekwente voegwoorde insluit.

'n Beduidende voordeel van die (semi-)outomatiese proses wat in hierdie artikel beskryf is, is die besparing aan mens-ure wat benut is in vergelyking met tyd wat normaalweg deur leksikograwe spandeer word aan die ontwikkeling van formele woordeboeke. Die konsepvoorstel wat geformuleer is, is nuttig vir die daarstel van meertalige aanlyn woordeboeke, saamgestel met behulp van gratis beskikbare aanlyn of elektroniese hulpbronne. Ongelukkig is selfs die beskikbaarstelling van elektroniese hulpmiddels, veral waar kleiner tale soos Afrikaans en Venda ter sprake is, ook nie altyd so volhoubaar nie. Een van die masjienvertaalsisteme wat ons in Julie 2021 vir die eksperiment gebruik het, english-afrikaans.co.za, was byvoorbeeld teen Januarie 2022 nie meer aanlyn beskikbaar nie. Webdienste soos hierdie se betroubaarheid word dan in twyfel getrek.

Die resulterende drietalige woordeboek wat saamgestel is, is reeds aanlyn as 'n *Lexonomy*-woordeboek beskikbaar om as 'n bewys van die konsep waarop verdere werk kan bou te dien[8]. Die feit dat die databasis wat die woordeboek onderlê in 'n masjienleesbare formaat, naamlik XML, afgelaai kan word, is belangrik vir indirekte teikengebruikers vir hergebruik om elektroniese hulpbronne te ontwikkel, veral vir hulpbronarm tale. Aangesien die navorsingspan ten gunste is van maksimale toegang tot elektroniese hulpbronne, veral vir die Afrikatale, stel ons ook die XML-weergawe van die woordeboek vir navorsers en ander gebruikers gratis beskikbaar op die webblad van SADiLaR (2022), 'n organisasie wat die skep van digitale hulpbronne vir die tale van Suid-Afrika ten doel het.

7.2 Toekomstige werk

'n Formele evaluering van die mate waarin die Afrikaanse vertalings akkurate

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1696 (Article)

vertaalekwivalente van die Venda lemmas weergee, sal veel bydra tot die waarde van die eksperiment. Hiervoor is die insette van moedertaalsprekers van Venda nodig. Dit sal verder ook interessant wees om die lemmalys van die woordeboek aan leksikografiese meetinstrumente soos dié van Prinsloo en De Schryver (2002) te toets ten einde vas te stel tot watter mate die verskillende alfabetiese strekke toereikend behandel is.

Die eksperiment soos hierbo beskryf maak dit moontlik om dié werkswyse na ander tale uit te brei en so binne 'n relatiewe kort tydperk en met heelwat minder mens-ure elektroniese woordeboeke bestaande uit verskillende taalpare beskikbaar te stel. Ten opsigte van die Afrikatale bestaan daar byvoorbeeld geen woordeboeke waarin beide die brontaal en die teikentaal Afrikatale is nie. Die metodologie soos hierbo beskryf maak die saamstel van sulke woordeboeke 'n haalbare onderneming.

Die woordeboek self kan verder uitgebrei word deur die datavelde uit te brei. Velde kan byvoorbeeld geskep word vir gebruiksvoorbeelde wat uit die korpus onttrek word en hoë-frekwensie kollokasies. Hierdie prosesse kan semi-outomaties met behulp van gratis korpusnavraagprogrammatuur uitgevoer word. Benewens die uitbreiding van die lemmalys soos deur frekwensie bepaal, is 'n kritiese evaluering van die huidige lemmalys nodig, met inagneming van die teikengebruiker. Verouderde of argaïese lemmas hoort waarskynlik nie in 'n aanleerderswoordeboek tuis nie en so 'n evaluering sal in samewerking met 'n Venda-spesialis gedoen moet word.

Eindnotas

1. Sien https://www.w3.org/standards/xml/core vir 'n volledige beskrywing van hierdie formaat.
2. http://autshumato.sourceforge.net/
3. https://en.wikipedia.org/wiki/Google_Translate
4. https://en.wikipedia.org/wiki/Microsoft_Translator
5. https://en.wikipedia.org/wiki/Yandex.Translate
6. https://wspel.wordpress.com/
7. https://wspel.wordpress.com/
8. 'n Gratis profiel kan op die Lexonomy-platform geregistreer word by https://www.lexonomy.eu/ en daarna is die konsepweergawe van die woordeboek by https://www.lexonomy.eu/POCVenEngAfr/ te sien.

Erkennings

— CBOLD

Die outeurs is dank en erkenning verskuldig aan die samestellers van die *Comparative Bantu OnLine Dictionary* (http://www.cbold.ish-lyon.cnrs.fr/) wat hulle woordeboekdata in elektroniese formaat beskikbaar stel vir nie-kommersiële gebruik deur ander navorsers.

— SADiLaR

Hierdie projek is moontlik gemaak met ondersteuning van SADiLaR (2022), 'n navorsingsinfrastruktuur wat deur die Departement van Wetenskap en Tegnologie van die Suid-Afrikaanse regering gestig is as deel van die Suid-Afrikaanse navorsingsinfrastruktuur-padkaart (SARIR).

— Nasionale Navorsingstigting (NNS)

Hierdie navorsing is finansiëel ondersteun deur die NNS. Die toekenninghouers (Unieke verwysings: S E Bosch (Toekenning nr. 109384) en E Taljard (Toekenning nr. 77735)) bevestig dat die menings, bevindinge en gevolgtrekkings of aanbevelings wat in enige NNS-ondersteunde navorsing uitgespreek word, hul eie is en dat die toekenningsinstansie geen aanspreeklikheid in dié verband aanvaar nie.

Bibliografie

Woordeboeke

Labuschagne, F.J. en L.C. Eksteen. 2010. *Pharos verklarende Afrikaanse woordeboek.* Kaapstad: Pharos Woordeboeke. Beskikbaar: https://www.pharosaanlyn.co.za/tuis

Murphy, M.L. 1997. *Venda: CBOLD (Comparative Bantu OnLine Dictionary).* Beskikbaar: http://www.cbold.ish-lyon.cnrs.fr/

Rundell, M. (Red.). 2002. *Macmillan English Dictionary for Advanced Learners*. Second edition. Oxford: Macmillan Education.

Van Warmelo, N.J. 1937. *Tshivenḓa–English Dictionary*. Pretoria: Staatsdrukker.

Wentzel, P.J. en T.W. Muloiwa. 1976. *Drietalige Elementêre Woordeboek / Trilingual Elementary Dictionary: Venda–Afrikaans–English.* Pretoria: Universiteit van Suid-Afrika.

Wentzel, P.J. en T.W. Muloiwa. 1982. *Ṱhalusamaipfi ya nyambotharu yo khwiniswaho: Luvenḓa–Luvhuru–Luisimane / Verbeterde drietalige woordeboek: Venḓa–Afrikaans–Engels / Improved Trilingual Dictionary: Venḓa–Afrikaans–English.* Pretoria: Universiteit van Suid-Afrika.

Ander bronne

Atkins, B.T.S., J. Clear en N. Ostler. 1992. Corpus Design Criteria. *Literary and Linguistic Computing* 7(1): 1-16.

Bartolomé-Díaz, B. en F. Frontini. 2020. Building a Domain-specific Bilingual Lexicon Resource with *Sketch Engine* and *Lexonomy*: Taking Ownership of the Issues. *Proceedings of the 2020 Globalex Workshop on Linked Lexicography, May 2020, Marseille, France*: 62-68. Marseille: European Language Resources Association. Beskikbaar: https://aclanthology.org/2020.globalex-1.11.pdf

CBOLD. 1997–2003. *Comparative Bantu OnLine Dictionary*. Beskikbaar: http://www.cbold.ish-lyon.cnrs.fr/.

Departement van Hoër Onderwys en Opleiding. 2015. National Qualifications Framework Act (67/2008): Revised Policy on the Minimum Requirements for Teacher Education Qualifications. *Government Gazette/Staatskoerant*, 19 Februarie 2015. Beskikbaar: https://bit.ly/31xp8rV

Eckart, T., S. Bosch, D. Goldhahn, U. Quasthoff en B. Klimek. 2019. Translation-based Dictionary Alignment for Under-resourced Bantu Languages. Eskevich, Maria, Gerard de Melo, Christian Fäth, John P. McCrae, Paul Buitelaar, Christian Chiarcos, Bettina Klimek en Milan Dojchinovski (Reds.). 2019. *2nd Conference on Language, Data and Knowledge (LDK 2019):* 17:1-17:11. Schloss Dagstuhl — Leibniz-Zentrum für Informatik: Dagstuhl Publishing. Beskikbaar: http://drops.dagstuhl.de/opus/volltexte/2019/10381/pdf/OASIcs-LDK-2019-17.pdf

European Lexicographic Infrastructure (ELEXIS). 2020. *Opening up Dictionaries, Linguistic Data and Language Tools for European Communities.* [Brochure]. Beskikbaar: https://elex.is/wp-content/uploads/2019/03/Print_Publicity_Brochure.pdf

Groves, M. en K. Mundt. 2015. Friend or Foe? Google Translate in Language for Academic Purposes. *English for Specific Purposes* 37: 112-121.

Jakubíček, M., M. Měchura, V. Kovář en P. Rychlý. 2018. Practical Post-Editing Lexicography with Lexonomy and Sketch Engine. 2018. *The XVIII EURALEX International Congress: Lexicography in Global Contexts, Ljubljana, Slovenia, July 17–21, 2018*. 65-67. Beskikbaar: https://euralex2018.cjvt.si/wp-content/uploads/sites/19/2020/08/Euralex2018_book_of_abstracts_FINAL.pdf

Jurafsky, D. en J.H. Martin. 2009. *Speech and Language Processing An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition.* New Jersey: Prentice Hall.

McKellar, C.A. en H.J. Groenewald. 2012. Frequency-based Data Selection for Statistical Machine Translation with Scarce Resources. Ndinga-Koumba-Binza, H.S. en S.E. Bosch (Reds.). 2012. *Language Science and Language Technology in Africa: A Festschrift for Justus C. Roux*: 271-290. Stellenbosch: SUN MeDIA.

Měchura, M. 2017. Introducing Lexonomy: An Open-source Dictionary Writing and Publishing System. Kosem, I. et al. (Hrsg.). 2017. *Electronic Lexicography in the 21st Century, Proceedings of eLex 2017 Conference, 19–21 September 2017, Leiden, the Netherlands:* 662-679. Brno: Lexical Computing CZ s.r.o.

Moors, C., I. Wilken, K. Calteaux en T. Gumede. 2018. Human Language Technology Audit 2018: Analysing the Development Trends in Resource Availability in all South African Languages. *Proceedings of the Annual Conference of the South African Institute of Computer Scientists and Information Technologists SAICSIT '18: Technology for Change, 26–28 September 2018, Port Elizabeth, South Africa:* 296-304. New York: The Association for Computing Machinery (ACM). Beskikbaar: https://doi.org/10.1145/3278681.3278716

Ṋemuṱamvuni, M.E. 2018. *Investigating the Effectiveness of Available Tools for Translating into Tshivenḓa.* M.A.-tesis, Universiteit van Suid-Afrika, Pretoria. Beskikbaar: http://hdl.handle.net/10500/25563

Poulos, G. 1990. *A Linguistic Analysis of Venda*. Pretoria: Via Afrika.

Prinsloo, D.J. en G.-M. de Schryver. 2002. Designing a Measurement Instrument for the Relative Length of Alphabetical Stretches in Dictionaries, with Special Reference to Afrikaans and English. Braasch, Anna and Claus Povlsen (Reds.). 2002. *Proceedings of the Tenth EURALEX*

International Congress, EURALEX 2002, Copenhagen, Denmark, August 13–17, 2002: 483-494. Copenhagen: Center for Sprogteknologi, Københavns Universitet.

SADiLaR. 2022. *South African Centre for Digital Language Resources.* Beskikbaar: https://sadilar.org/index.php/en/

Sager, J.C. 1994. *Language Engineering and Translation: Consequences of Automation.* Amsterdam/ Philadelphia: John Benjamins.

Stemle, E.W., A. Abel en V. Lyding. 2019. Language Varieties Meet One-Click Dictionary. Kosem, I. et al. (Reds.). 2019. Electronic Lexicography in the 21st Century: Smart Lexicography. Proceedings of the eLex 2019 *Conference, 1–3 October* 2019*, Sintra, Portugal*: 537-546. Brno, Czech Republic: Lexical Computing CZ s.r.o. Beskikbaar: https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_31.pdf

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1697 (Article)

Information Needs and Contextualization in the Consultation Process of Dictionaries that are Linked to e-Texts

Theo J.D. Bothma, *Department of Information Science, University of Pretoria, Pretoria, South Africa (theo.bothma@up.ac.za)*
and
Rufus H. Gouws, *Department of Afrikaans and Dutch, Stellenbosch University, Stellenbosch, South Africa (rhg@sun.ac.za)*

Abstract: This article focuses on various aspects regarding contextualization when e-texts are linked to integrated dictionaries. The article responds to a twofold problem statement: (1) Dictionaries linked to e-texts do not sufficiently take into account the contextualization and cotextualization of words when providing information to users. (2) The integrated dictionary may contain the items needed for contextualization and cotextualization, but the e-device cannot interpret the context of a word and link the word to the relevant item in the dictionary article. The aim of the article is to show the need of linking a word from a text on an e-device to the correct sense in the integrated dictionary. This presupposes dynamic dictionary articles and lexicographic structures in which a relation between words in an e-text and user-specified lexicographic sources is established. Some existing projects that perform such linking are discussed and evaluated. Based on these results this article makes some suggestions. It is foreseen that there will be a "black box" of software between the selected word and the dictionary that will determine the correct lemma and sense to be selected from the e-dictionary. Having discussed various alternatives, the article suggests parallel contextualization between the dictionary and the software of the e-device. Many aspects discussed in this article require further research. Relevant proposals are made with regard to this research.

Keywords: CONTEXT, CONTEXTUALIZATION, COTEXT, DICTIONARY CONSULTATION, E-DEVICE, E-READER, E-TEXT, INTEGRATED DICTIONARY, LEXICOGRAPHIC NEEDS, LINKING, PARALLEL CONTEXTUALIZATION, TEXT RECEPTION

Opsomming: Inligtingsbehoeftes en kontekstualisering in die raadpleging van woordeboeke wat aan e-tekste gekoppel is. Hierdie artikel fokus op verskeie aspekte van kontekstualisering wanneer e-tekste gekoppel word aan geïntegreerde woordeboeke. Die artikel het 'n tweevoudige probleemstelling: (1) Woordeboeke wat aan e-tekste gekoppel is, verreken nie die kontekstualisering en kotekstualisering van woorde genoegsaam wanneer inligting aan gebruikers gebied word nie. (2) Die geïntegreerde woordeboek mag wel die aanduiders

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1697 (Article)

wat nodig is vir kontekstualisering en kotekstualisering bevat, maar die e-apparaat kan nie die konteks van 'n woord interpreteer en dit aan die tersaaklike aanduider in die woordeboek koppel nie. Die doel van hierdie artikel is om die behoefte aan te toon om 'n woord in 'n teks op 'n e-apparaat aan die regte betekenisonderskeiding in die geïntegreerde woordeboek te koppel. Dit vooronderstel dinamiese woordeboekartikels en leksikografiese strukture waarin 'n verhouding tussen woorde in 'n e-teks en gebruikerbepaalde leksikografiese bronne gevestig word. Enkele bestaande projekte waarin hierdie soort koppeling voorkom, word bespreek en geëvalueer. Na aanleiding van die resultate hiervan word bepaalde voorstelle gemaak. Dit word voorsien dat daar 'n "black box" met sagteware tussen die gekose woord en die woordeboek sal wees wat die korrekte lemma en betekenisonderskeiding in die e-woordeboek sal bepaal. Verskeie alternatiewe word bespreek waarna parallelle kontekstualisering tussen die woordeboek en die sagteware van die e-apparaat voorgestel word. Baie aspekte wat in hierdie artikel bespreek word, vereis verdere navorsing en relevante voorstelle word in hierdie verband gemaak.

Sleutelwoorde: E-APPARAAT, E-LESER, E-TEKS, GEÏNTEGREERDE WOORDEBOEK, KONTEKS, KONTEKSTUALISERING, KOTEKS, KOPPELING, LEKSIKOGRAFIESE BEHOEFTES, PARALLELLE KONTEKSTUALISERING, TEKSBEGRIP, WOORDEBOEKRAADPLEGING

1. Introduction

Lexicographic needs arise in extra-lexicographical situations and such needs initiate the execution of dictionary consultation procedures. A dictionary user finds a word in a text that causes text reception problems, and he/she consults a dictionary to find that word and the appropriate guidance to solve the problem.

Dictionary users typically do not read a dictionary, but they consult a dictionary for immediate needs, for example, to retrieve a limited amount of information to solve a specific problem (Tarp 2012). Successful dictionary consultation is achieved when these punctual needs can be satisfied because the information that had to be retrieved falls within the scope of the genuine purpose of the specific dictionary. The genuine purpose of a dictionary, according to Wiegand (1998: 299), lies therein that it can be used to retrieve specific information from the lexicographic data accommodated in the partial texts with outer access structure, regarding certain features of those linguistic expressions that belong to the subject matter of the dictionary. Achieving the genuine purpose of a dictionary and satisfying a punctual need are often impeded not by the data available in the dictionary articles but by the lack of supporting items to ensure an optimal retrieval of information. This is because dictionary articles contain a sufficient variety of items that convey the relevant lexicographic data but an insufficient number of contextual and cotextual items to supplement the other items.

A dictionary reflects the lexicon of the specific language by means of the lemma selection that enables a representative macrostructural coverage. However, in a dictionary a lemma sign as guiding element of an article is isolated from its occurrence in the real language. Sufficient addressing procedures are

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1697 (Article)

required to counter this lexicographic isolation.

In the structuring of their dictionaries and the way in which data are presented, lexicographers need to take cognizance not only of the lexicographic needs of their intended target users but also of their reference skills. Although it is often required from the users to apply their mind when consulting a dictionary (i.e. carefully evaluate the search results to ensure that the suggestion by the system is correct and/or acceptable in the given context), a user-friendly approach is needed because the default presentation in many dictionaries does not guarantee consultation success. Retrieving information from the lexicographic data is often further impeded by the density of dictionary articles, unnatural syntax and data overload. In the online environment, a specific dictionary is often linked to a specific device, for example an e-reader. The user is guided from a word found in a text on the e-reader to the treatment of that word in the integrated dictionary. However, such a consultation procedure often fails because the user cannot retrieve the required information due to the linking not being directed at the appropriate item in the dictionary article, or even to an incorrect dictionary article. This could be because the dictionary does not offer enough contextual and cotextual assistance or because the software of the e-reader cannot identify the appropriate context in the text or link it to the appropriate item in the dictionary.

This leads to the following twofold problem statement to which this article will respond:

— Dictionaries linked to e-devices do not sufficiently take into account the contextualization and cotextualization of words when providing information to users.
— The integrated dictionary may contain the items needed for contextualization and cotextualization, but the e-reader cannot interpret the context of a word and link the word to the relevant item in the dictionary article.

Although some traditional procedures can be maintained to provide a certain degree of contextualization and cotextualization, lexicography in a new era is in need of new procedures. One such relatively new procedure is to integrate writing assistants in dictionaries with a text production function. This paper focuses primarily on text reception needs and the use of linking procedures between an e-device and the integrated dictionary to enhance contextualization and cotextualization. A point of departure is that lexicographers should be aware of the typical occurrence of words in real texts and the lexicographic process prevailing in integrated products should enable the recontextualization of these words.

2. A traditional approach to context and cotext in dictionaries and a wider use of the terms

When discussing a topic like contextualization and dictionaries, it is important to

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1697 (Article)

have a clear understanding of the use of the relevant terms in the field of metalexicography and the lexicographic practice. In metalexicography, cf. Wiegand (1988), Gouws (2002), Gouws and Prinsloo (2005), Lettner (2020), Domínguez and Gouws (in print), a distinction is made between items giving the context and those giving the cotext in dictionaries. Context is regarded as the pragmatic environment of an item and is indicated in dictionaries by, for example, labels, glosses and cultural notes. Cotext refers to the textual environment of an item and is typically indicated in a dictionary article by means of example sentences and collocations. The position allocated to items giving the cotext is determined by the type of microstructure of a specific dictionary. If the dictionary has an integrated microstructure, the cotextual items are given in the same subcomment on semantics where the relevant translation equivalent or paraphrase of meaning is given. This leads to a process of direct non-lemmatic addressing. If the dictionary has a non-integrated microstructure, the cotext items are presented in a separate text block but with a clear indication of which cotext item belongs to which translation equivalent or paraphrase of meaning. Remote addressing prevails in such a dictionary article.

In this contribution, the lexicographical use of the terms context and cotext will be maintained. However, because the dictionaries discussed in this paper are not used in isolation but always as part of an integrated product with an e-device as the other component, a slightly wider use of these terms will be proposed. They will have a more comprehensive scope than a mere use in dictionaries. The context of a word is therefore also regarded as its occurrence in a dictionary-external text, for example in a text downloaded onto an e-reader or viewed in a browser on any electronic device. Here the context includes the source and specific volume of the text, the chapter, section, and paragraph where the word occurs as well as extra-textual information regarding the author of the text and the period and geographical environment where the text is situated. The cotext of the word remains its syntactic environment, but, besides its occurrence in a sentence, also the paragraph in which it occurs.

As indicated earlier in this article, the immediate need that leads to a dictionary consultation originates in an extra-lexicographic situation. In this paper the extra-lexicographic environment where the need originates, will be a specific text downloaded on an e-reader or viewed in a browser on any electronic device. The user is guided from a word in such a dictionary-external context to the word presented as lemma sign in a specific dictionary integrated with the e-device. The e-device may contain more than one dictionary. In the selected dictionary the items giving context and cotext should enable the user to link a specific treatment of the word in the dictionary to a specific occurrence of that word in the extra-lexicographic environment.

The focus of this article is to negotiate the contextualization and cotextualization of words as they occur in texts to improve the satisfaction of dictionary users with regard to especially their text reception information needs. This kind of contextualization implies that lexicographers should be acutely aware of the typical dictionary-internal contexts, and they should be able to relate dictionary-external words to these items. When developing the software associated with an e-reader or other e-device, one has to be aware of the extent of contextualization in the linked dictionary. The software needs to be adapted to identify the context of a word and to use that to ensure a successful linking to an item in the dictionary.

3. Dictionary-external context

Depending on their functions, dictionaries should include ample items to supply the appropriate contextual and cotextual guidance to their users. In a dictionary article the word represented by the lemma sign is treated in isolation. The contextual and cotextual items provided as part of the lexicographic treatment should not be selected in a haphazard way but, utilising a balanced and representative corpus, it should reflect something of the typical dictionary-external occurrence of the word. This should enable the user to link the word as it was encountered in an extra-lexicographic environment (and context) to a specific search zone in the dictionary that contains the relevant treatment of that word. The context and cotext from dictionary-external occurrences of the word should be transferred to the dictionary article to enhance text reception and text production procedures.

Employing a more comprehensive use of the terms *context* and *cotext* (and also *contextualization* and *cotextualization*), contextualization should not only be seen as referring to a dictionary-internal procedure. Within a dictionary, lexicographers focus on giving the context of the treated word. However, another context and another procedure of contextualization should also be recognized by lexicographers. This is the context outside the dictionary, in this paper the texts found on an e-device. This will determine the dictionary or dictionaries to be integrated with the e-device. Contextualization then also implies an anchoring between this dictionary-external context and items presented in dictionary articles, and it determines the way in which the relevant data are negotiated in the dictionary-internal ordering and presentation procedures.

The online environment offers different possibilities of satisfying lexicographic information needs by means of contextualization. Dictionaries still function as stand-alone products or they can be part of a dictionary portal (Engelberg and Müller-Spitzer 2013: 1023). In both these instances the contextualization in the dictionary is not motivated by specific texts in the dictionary-external environment. The editorial system of the dictionary requires that certain items in the dictionary article, for example, paraphrases of meaning or translation equivalents, should be addressed by items giving context and cotext, and these supporting data are either made-up by the lexicographer or extracted from the specific corpus used by the lexicographer for the specific dictionary. Context can be obtained beyond the stand-alone dictionary of the dictionary portal. The online environment enables such possibilities.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1697 (Article)

Where dictionary users get access from within a dictionary, a search region, or a dictionary portal, a search domain, to the internet and other sources outside the dictionary portal, a search universe (cf. Gouws 2021: 7), a comprehensive but often unspecified and uncurated pool of supporting data is at the disposal of the user. This offers an opportunity to link an item in a dictionary to a dictionary-external source, but the users must apply their mind to make the appropriate pairing. The search moves from the dictionary to the external source to satisfy a specific lexicographic need of a user. For the current paper, the focus is on the reverse search direction — and this is also possible in an online environment. Users of an e-device should have the opportunity to link an item in a dictionary-external source to a lemma sign in a specific dictionary and the items in a specific subcomment on semantics in the article of that lemma sign. The ideal is that the pairing will link the word in the dictionary-external text to the appropriate items in the dictionary article so that a user can achieve an unambiguous retrieval of information.

To achieve the above-mentioned consultation, lexicographers need to take cognizance of another type of contextualization procedure. The e-device and its software should be able to identify the context of a word in the text and link this context with the appropriate context in the integrated dictionary. This type of contextualization by means of linking is discussed in the next section.

4. Linking

Linking forms the basis for establishing a mapping between a word in a text and an item in an e-dictionary article. Linking as a contextual procedure presupposes dynamic dictionary articles and lexicographic structures in which a relation between words in an e-text and user-specified lexicographic sources is established. A problematic word encountered in an extra-lexicographic environment needs to be linked to the treatment of that word and its specific sense in a dictionary. Successful linking would map the contextualization/cotextualization of a word in the source to contextualization/cotextualization of a lemma in the dictionary which would result in users being linked to exact and relevant items. It therefore needs both texts and dictionary articles with higher contextualization potential.

In the remainder of this section the focus will be limited to systems that make use of linking, and to context and cotext in dictionaries.

5. Selected projects that link e-texts to e-dictionaries

Linking words in an e-text to language tools, especially e-dictionaries, is not a new concept, and has been implemented in various projects. The following projects are discussed and briefly evaluated:

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1697 (Article)

- The Perseus Project
- Amazon Kindle dictionary linking
- Browser-based linking
- Linking in an e-learning environment

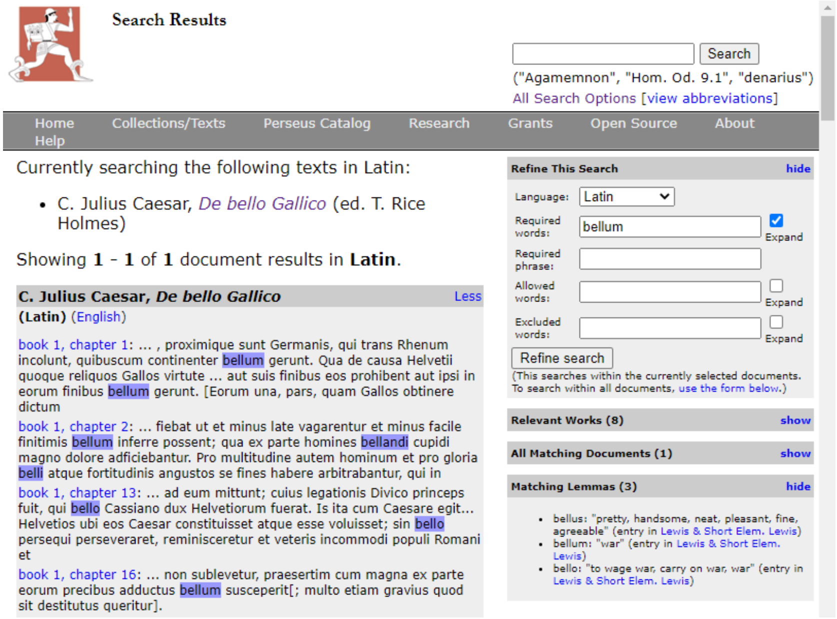
In each case the project is briefly outlined and examples are discussed. Following these discussions, the principles involved in the projects are briefly evaluated.

5.1 Perseus Project

The Perseus Digital Library (http://www.perseus.tufts.edu/hopper/) is a project that explores the possibilities that online digital collections offer. The project "covers the history, literature and culture of the Greco-Roman world" (Perseus Digital Library, n.d.-b), and, since its inception in 1987, expanded to include "a massive library of art objects, sites, and buildings", Arabic, Germanic, 19th-Century American Materials etc. (Perseus Digital Library, n.d.-c). According to Crane (1998) the "long-term goal must be to make accessible, both physically and intellectually, to every human being on this planet the complete record of humanity".

The Greek and Roman collections currently contain 44,462,693 English words, 13,507,448 Greek words and 10,525,338 Latin words. The Greek and Latin texts are all encoded with TEI (Perseus Digital Library, n.d.-a, Rydberg-Cox et al. 2000) to provide easy access to properties of individual words so that they can be studied in depth. The encoding allows the user to search for a lemma, and obtain all inflected forms related to the lemma, either in all the texts, or in a specified subset. For example, in Figure 1, the word "*bellum*" is searched in the *De Bello Gallico* by Julius Caesar, which results in the highlighted words in the text; at the bottom right of the image, the three possible lemmas are given, viz. "*bellus*", "*bellum*" and "*bello*" By clicking on a specific occurrence, in this case "*bello*", all possible parts of speech are given, as illustrated in Figure 2. The Latin Word Study Tool (Figure 3) provides a statistical probability of all possible correct part of speech (PoS) analyses, and selects one of the options as the most likely one in context, but adds a caveat: "It may or may not be the correct form." It also provides a link to two online Latin dictionaries, viz. Lewis and Short (see Figure 3 for a short extract, which offers the meaning "war" as translation option) and Elementary Lewis and Short. This recommendation is correct, but this is unfortunately not always the case; if the PoS parsing statistical recommendation were to be incorrect, translations of "pretty, handsome" or "to wage war" would be possible, as is evident form the three possible lemmas listed in Figure 1, and with the PoS analyses in Figures 2, 3 and 4.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1697 (Article)



Search Results

Search
("Agamemnon", "Hom. Od. 9.1", "denarius")
All Search Options [view abbreviations]

Currently searching the following texts in Latin:

- C. Julius Caesar, *De bello Gallico* (ed. T. Rice Holmes)

Showing **1** - **1** of **1** document results in **Latin**.

C. Julius Caesar, *De bello Gallico* Less
(Latin) (English)

book 1, chapter 1: ... , proximique sunt Germanis, qui trans Rhenum incolunt, quibuscum continenter bellum gerunt. Qua de causa Helvetii quoque reliquos Gallos virtute ... aut suis finibus eos prohibent aut ipsi in eorum finibus bellum gerunt. [Eorum una, pars, quam Gallos obtinere dictum

book 1, chapter 2: ... fiebat ut et minus late vagarentur et minus facile finitimis bellum inferre possent; qua ex parte homines bellandi cupidi magno dolore adficiebantur. Pro multitudine autem hominum et pro gloria belli atque fortitudinis angustos se fines habere arbitrabantur, qui in

book 1, chapter 13: ... ad eum mittunt; cuius legationis Divico princeps fuit, qui bello Cassiano dux Helvetiorum fuerat. Is ita cum Caesare egit... Helvetios ubi eos Caesar constituisset atque esse voluisset; sin bello persequi perseveraret, reminisceretur et veteris incommodi populi Romani et

book 1, chapter 16: ... non sublevetur, praesertim cum magna ex parte eorum precibus adductus bellum susceperit[; multo etiam gravius quod sit destitutus queritur].

Refine This Search hide

Language: Latin
Required words: bellum ☑ Expand
Required phrase:
Allowed words: □ Expand
Excluded words: □ Expand
Refine search
(This searches within the currently selected documents. To search within all documents, use the form below.)

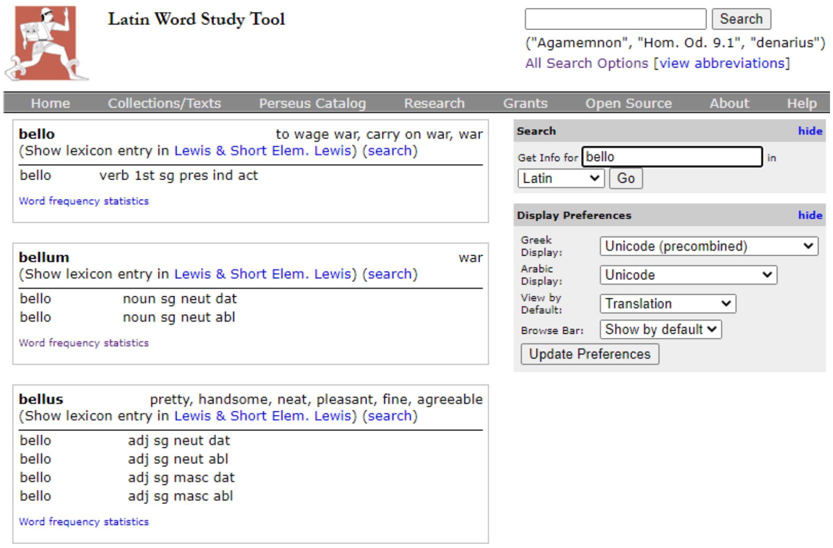
Relevant Works (8) show

All Matching Documents (1) show

Matching Lemmas (3) hide

- bellus: "pretty, handsome, neat, pleasant, fine, agreeable" (entry in Lewis & Short Elem. Lewis)
- bellum: "war" (entry in Lewis & Short Elem. Lewis)
- bello: "to wage war, carry on war, war" (entry in Lewis & Short Elem. Lewis)

Figure 1: The search word is "*bellum*", and all potential derivatives in the specific text are found; the selection lists "*bellum*", "*bellandi*", "*belli*" and "*bello*"



Latin Word Study Tool

Search
("Agamemnon", "Hom. Od. 9.1", "denarius")
All Search Options [view abbreviations]

bello to wage war, carry on war, war
(Show lexicon entry in Lewis & Short Elem. Lewis) (search)

bello	verb 1st sg pres ind act

Word frequency statistics

bellum war
(Show lexicon entry in Lewis & Short Elem. Lewis) (search)

bello	noun sg neut dat
bello	noun sg neut abl

Word frequency statistics

bellus pretty, handsome, neat, pleasant, fine, agreeable
(Show lexicon entry in Lewis & Short Elem. Lewis) (search)

bello	adj sg neut dat
bello	adj sg neut abl
bello	adj sg masc dat
bello	adj sg masc abl

Word frequency statistics

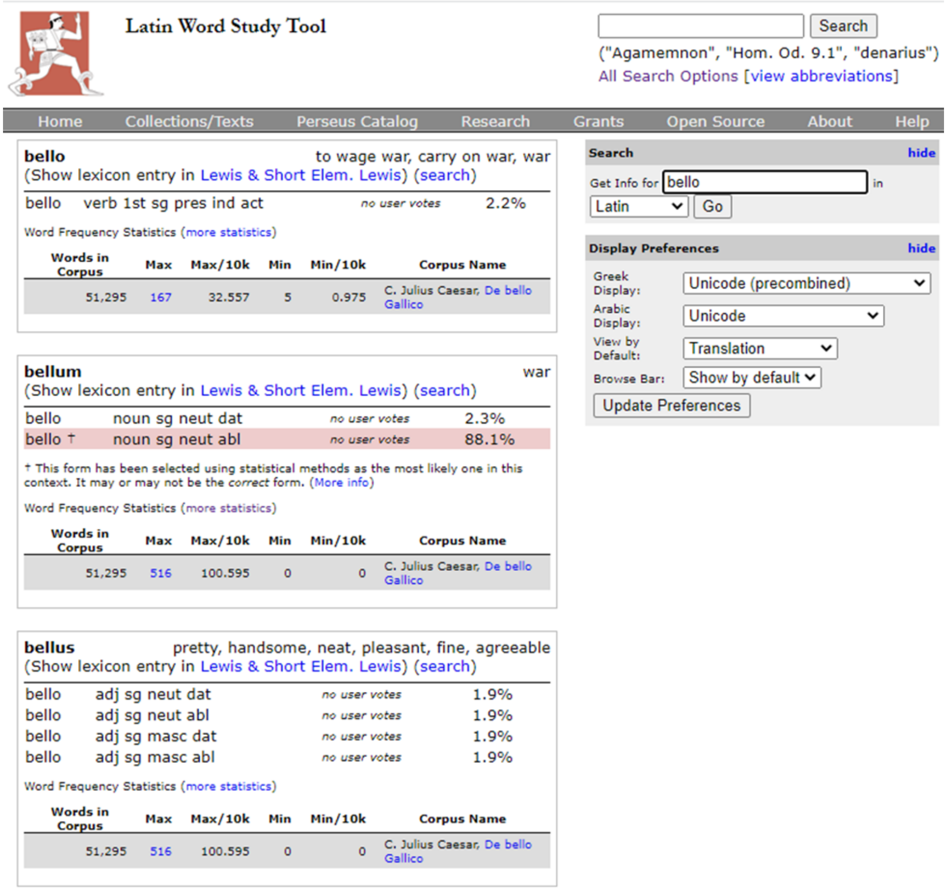
Search hide

Get Info for bello in Latin Go

Display Preferences hide

Greek Display: Unicode (precombined)
Arabic Display: Unicode
View by Default: Translation
Browse Bar: Show by default
Update Preferences

Figure 2: Potential part-of-speech analyses of the selected item, "*bello*"



bello to wage war, carry on war, war
(Show lexicon entry in Lewis & Short Elem. Lewis) (search)

bello	verb 1st sg pres ind act	*no user votes*	2.2%

Word Frequency Statistics (more statistics)

Words in Corpus	Max	Max/10k	Min	Min/10k	Corpus Name
51,295	167	32.557	5	0.975	C. Julius Caesar, De bello Gallico

bellum war
(Show lexicon entry in Lewis & Short Elem. Lewis) (search)

bello	noun sg neut dat	*no user votes*	2.3%
bello †	noun sg neut abl	*no user votes*	88.1%

† This form has been selected using statistical methods as the most likely one in this context. It may or may not be the *correct* form. (More info)

Word Frequency Statistics (more statistics)

Words in Corpus	Max	Max/10k	Min	Min/10k	Corpus Name
51,295	516	100.595	0	0	C. Julius Caesar, De bello Gallico

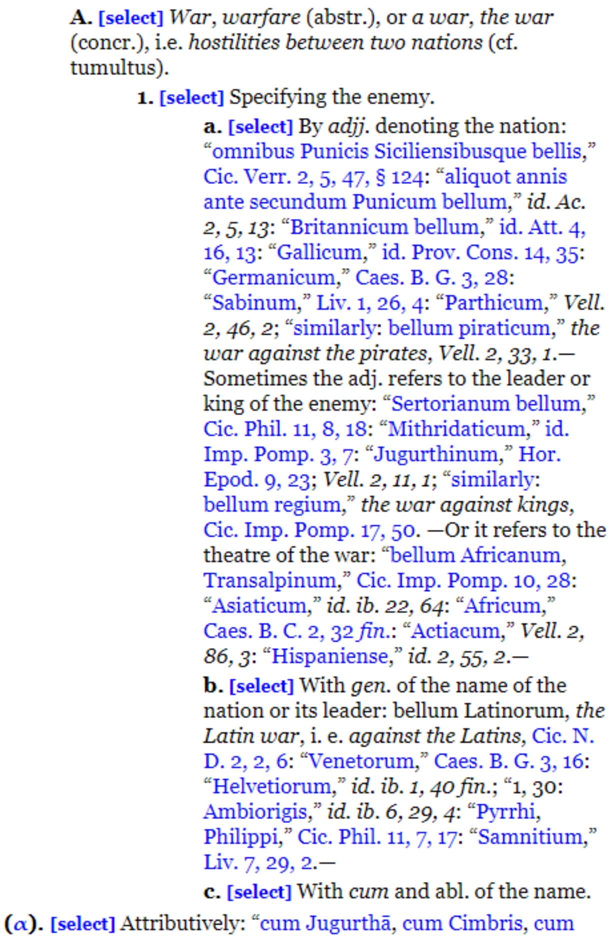
bellus pretty, handsome, neat, pleasant, fine, agreeable
(Show lexicon entry in Lewis & Short Elem. Lewis) (search)

bello	adj sg neut dat	*no user votes*	1.9%
bello	adj sg neut abl	*no user votes*	1.9%
bello	adj sg masc dat	*no user votes*	1.9%
bello	adj sg masc abl	*no user votes*	1.9%

Word Frequency Statistics (more statistics)

Words in Corpus	Max	Max/10k	Min	Min/10k	Corpus Name
51,295	516	100.595	0	0	C. Julius Caesar, De bello Gallico

Figure 3: Statistical analysis suggests that "noun sg neut abl" is correct, and provides the meaning "war"



A. [select] *War, warfare* (abstr.), or *a war, the war* (concr.), i.e. *hostilities between two nations* (cf. tumultus).

1. [select] Specifying the enemy.

a. [select] By *adjj.* denoting the nation: "omnibus Punicis Siciliensibusque bellis," Cic. Verr. 2, 5, 47, § 124: "aliquot annis ante secundum Punicum bellum," *id. Ac. 2, 5, 13*: "Britannicum bellum," id. Att. 4, 16, 13: "Gallicum," id. Prov. Cons. 14, 35: "Germanicum," Caes. B. G. 3, 28: "Sabinum," Liv. 1, 26, 4: "Parthicum," *Vell. 2, 46, 2*; "similarly: bellum piraticum," *the war against the pirates, Vell. 2, 33, 1*.—Sometimes the adj. refers to the leader or king of the enemy: "Sertorianum bellum," Cic. Phil. 11, 8, 18: "Mithridaticum," id. Imp. Pomp. 3, 7: "Jugurthinum," Hor. Epod. 9, 23; *Vell. 2, 11, 1*; "similarly: bellum regium," *the war against kings*, Cic. Imp. Pomp. 17, 50. —Or it refers to the theatre of the war: "bellum Africanum, Transalpinum," Cic. Imp. Pomp. 10, 28: "Asiaticum," *id. ib. 22, 64*: "Africum," Caes. B. C. 2, 32 *fin.*: "Actiacum," *Vell. 2, 86, 3*: "Hispaniense," *id. 2, 55, 2*.—

b. [select] With *gen.* of the name of the nation or its leader: bellum Latinorum, *the Latin war*, i. e. *against the Latins*, Cic. N. D. 2, 2, 6: "Venetorum," Caes. B. G. 3, 16: "Helvetiorum," *id. ib. 1, 40 fin.*; "1, 30: Ambiorigis," *id. ib. 6, 29, 4*: "Pyrrhi, Philippi," Cic. Phil. 11, 7, 17: "Samnitium," Liv. 7, 29, 2.—

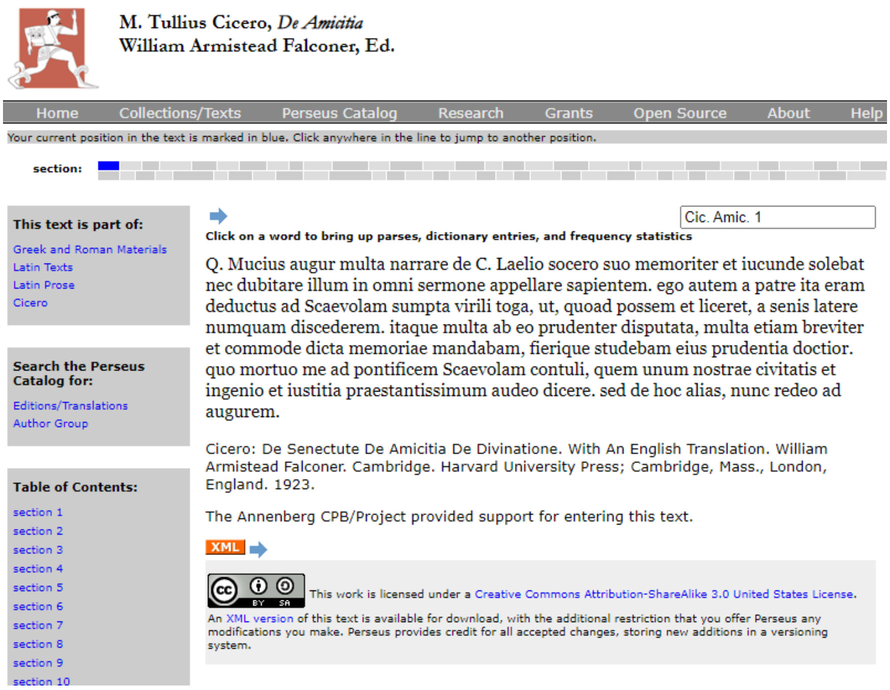
c. [select] With *cum* and abl. of the name.

(α). [select] Attributively: "cum Jugurthā, cum Cimbris, cum

Figure 4: An extract from the linked version of the Latin–English dictionary by Lewis and Short in the Perseus Project, which provides more detailed information about "*bellum*" and links to various texts

The following figures provide examples of incorrect morphological parsing and/or the complexity of finding the correct translation equivalent. The text, from the *De Amicitia* by Cicero, is given in Figure 5.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1697 (Article)



M. Tullius Cicero, *De Amicitia*
William Armistead Falconer, Ed.

Your current position in the text is marked in blue. Click anywhere in the line to jump to another position.

This text is part of:

Greek and Roman Materials
Latin Texts
Latin Prose
Cicero

Search the Perseus Catalog for:

Editions/Translations
Author Group

Table of Contents:

section 1
section 2
section 3
section 4
section 5
section 6
section 7
section 8
section 9
section 10

Click on a word to bring up parses, dictionary entries, and frequency statistics

Q. Mucius augur multa narrare de C. Laelio socero suo memoriter et iucunde solebat nec dubitare illum in omni sermone appellare sapientem. ego autem a patre ita eram deductus ad Scaevolam sumpta virili toga, ut, quoad possem et liceret, a senis latere numquam discederem. itaque multa ab eo prudenter disputata, multa etiam breviter et commode dicta memoriae mandabam, fierique studebam eius prudentia doctior. quo mortuo me ad pontificem Scaevolam contuli, quem unum nostrae civitatis et ingenio et iustitia praestantissimum audeo dicere. sed de hoc alias, nunc redeo ad augurem.

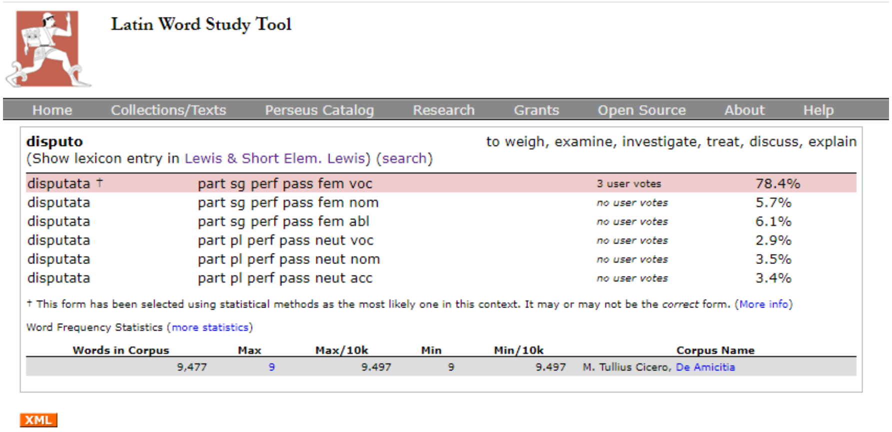
Cicero: De Senectute De Amicitia De Divinatione. With An English Translation. William Armistead Falconer. Cambridge. Harvard University Press; Cambridge, Mass., London, England. 1923.

The Annenberg CPB/Project provided support for entering this text.

This work is licensed under a Creative Commons Attribution-ShareAlike 3.0 United States License.

An XML version of this text is available for download, with the additional restriction that you offer Perseus any modifications you make. Perseus provides credit for all accepted changes, storing new additions in a versioning system.

Figure 5: Three words in line 4 are relevant — "*ab*", "*eo*" and "*disputata*" (discussed in inverse order)



Latin Word Study Tool

disputo to weigh, examine, investigate, treat, discuss, explain
(Show lexicon entry in Lewis & Short Elem. Lewis) (search)

disputata †	part sg perf pass fem voc	3 user votes	78.4%
disputata	part sg perf pass fem nom	*no user votes*	5.7%
disputata	part sg perf pass fem abl	*no user votes*	6.1%
disputata	part pl perf pass neut voc	*no user votes*	2.9%
disputata	part pl perf pass neut nom	*no user votes*	3.5%
disputata	part pl perf pass neut acc	*no user votes*	3.4%

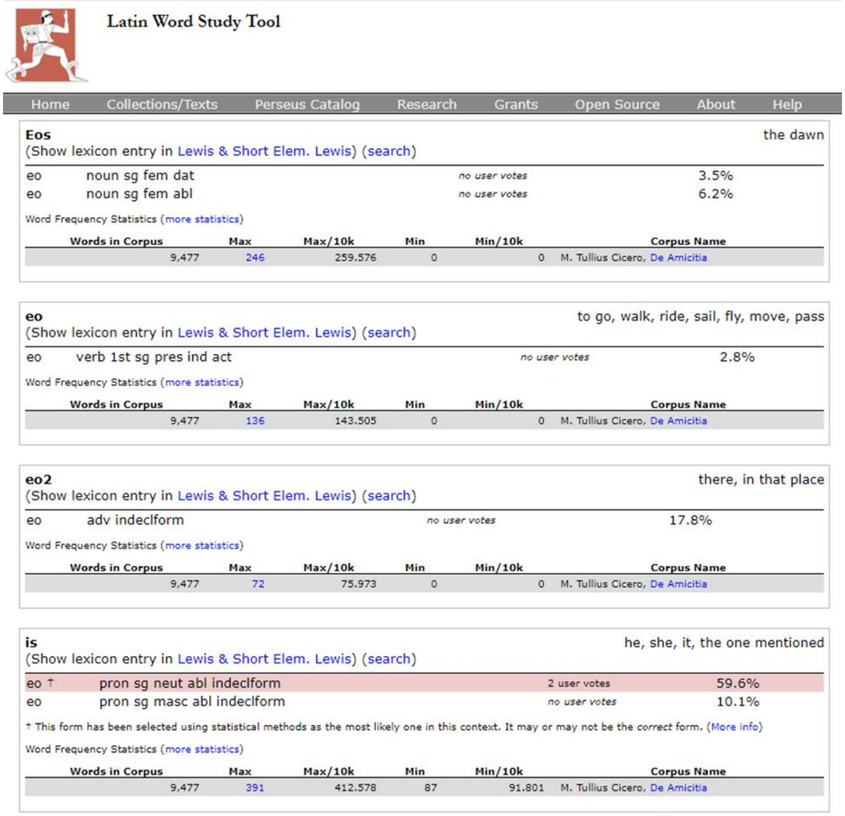
† This form has been selected using statistical methods as the most likely one in this context. It may or may not be the *correct* form. (More info)

Word Frequency Statistics (more statistics)

Words in Corpus	Max	Max/10k	Min	Min/10k	Corpus Name
9,477	9	9.497	9	9.497	M. Tullius Cicero, De Amicitia

Figure 6: The lemma selection for "*disputata*" is correct. The PoS is partially correct, as it is not "fem voc", as suggested by the statistical analysis, but "neut acc"

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1697 (Article)



Eos — the dawn
(Show lexicon entry in Lewis & Short Elem. Lewis) (search)

eo	noun sg fem dat	no user votes	3.5%
eo	noun sg fem abl	no user votes	6.2%

Word Frequency Statistics (more statistics)

Words in Corpus	Max	Max/10k	Min	Min/10k	Corpus Name
9,477	246	259.576	0	0	M. Tullius Cicero, De Amicitia

eo — to go, walk, ride, sail, fly, move, pass
(Show lexicon entry in Lewis & Short Elem. Lewis) (search)

eo	verb 1st sg pres ind act	no user votes	2.8%

Word Frequency Statistics (more statistics)

Words in Corpus	Max	Max/10k	Min	Min/10k	Corpus Name
9,477	136	143.505	0	0	M. Tullius Cicero, De Amicitia

eo2 — there, in that place
(Show lexicon entry in Lewis & Short Elem. Lewis) (search)

eo	adv indeclform	no user votes	17.8%

Word Frequency Statistics (more statistics)

Words in Corpus	Max	Max/10k	Min	Min/10k	Corpus Name
9,477	72	75.973	0	0	M. Tullius Cicero, De Amicitia

is — he, she, it, the one mentioned
(Show lexicon entry in Lewis & Short Elem. Lewis) (search)

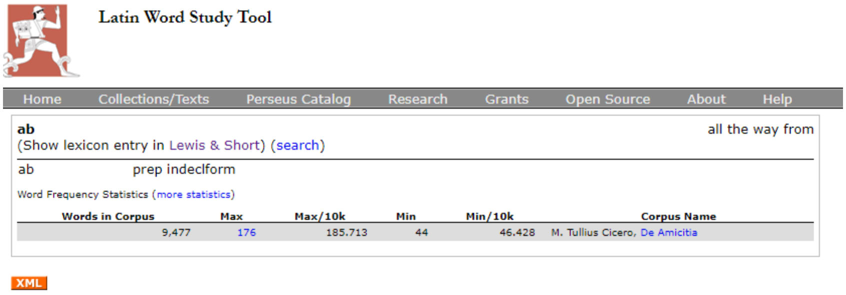
eo †	pron sg neut abl indeclform	2 user votes	59.6%
eo	pron sg masc abl indeclform	no user votes	10.1%

† This form has been selected using statistical methods as the most likely one in this context. It may or may not be the *correct* form. (More info)

Word Frequency Statistics (more statistics)

Words in Corpus	Max	Max/10k	Min	Min/10k	Corpus Name
9,477	391	412.578	87	91.801	M. Tullius Cicero, De Amicitia

Figure 7: There are four possible lemmas for "*eo*"; the statistical analysis identifies the correct lemma, "*eo*", but the incorrect gender, as it is "masc", and not "neut"



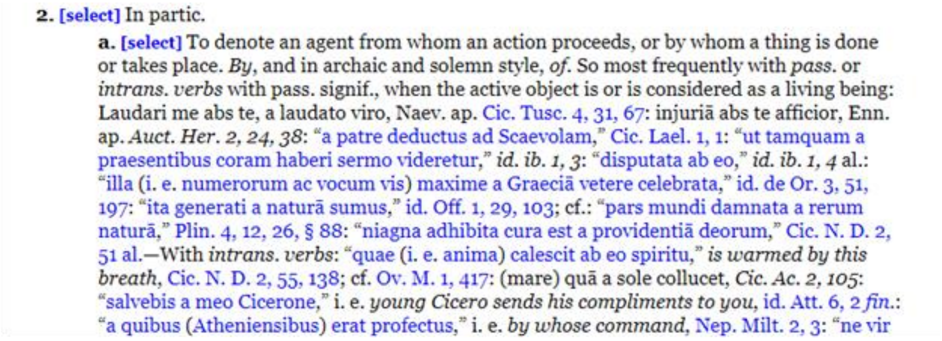
ab — all the way from
(Show lexicon entry in Lewis & Short) (search)

ab	prep indeclform

Word Frequency Statistics (more statistics)

Words in Corpus	Max	Max/10k	Min	Min/10k	Corpus Name
9,477	176	185.713	44	46.428	M. Tullius Cicero, De Amicitia

Figure 8: The preposition "*ab*" is identified correctly; however, one translation equivalent is provided, "all the way from", which makes no sense in context



2. [select] In partic.

a. [select] To denote an agent from whom an action proceeds, or by whom a thing is done or takes place. *By*, and in archaic and solemn style, *of*. So most frequently with *pass.* or *intrans. verbs* with pass. signif., when the active object is or is considered as a living being: Laudari me abs te, a laudato viro, Naev. ap. Cic. Tusc. 4, 31, 67: injuriā abs te afficior, Enn. ap. *Auct. Her.* 2, 24, 38: "a patre deductus ad Scaevolam," Cic. Lael. 1, 1: "ut tamquam a praesentibus coram haberi sermo videretur," *id. ib.* 1, 3: "disputata ab eo," *id. ib.* 1, 4 al.: "illa (i. e. numerorum ac vocum vis) maxime a Graeciā vetere celebrata," id. de Or. 3, 51, 197: "ita generati a naturā sumus," id. Off. 1, 29, 103; cf.: "pars mundi damnata a rerum naturā," Plin. 4, 12, 26, § 88: "niagna adhibita cura est a providentiā deorum," Cic. N. D. 2, 51 al.—With *intrans. verbs*: "quae (i. e. anima) calescit ab eo spiritu," *is warmed by this breath*, Cic. N. D. 2, 55, 138; cf. Ov. M. 1, 417: (mare) quā a sole collucet, *Cic. Ac.* 2, 105: "salvebis a meo Cicerone," i. e. *young Cicero sends his compliments to you*, id. Att. 6, 2 *fin.*: "a quibus (Atheniensibus) erat profectus," i. e. *by whose command*, Nep. Milt. 2, 3: "ne vir

Figure 9: Clicking on "Show lexicon entry in Lewis and Short" provides the full entry for "*ab*" in the dictionary — a short selection of the correct sense is provided

The entry in Lewis and Short, however, provides a total overload of information: the article is more than 5,800 words long, and the correct translation equivalent occurs in a similar text (also by Cicero) in the hierarchy at II.B.2.a, at around word 3,400.

The preceding discussion is not intended to be a full evaluation of the Perseus project features and functionalities. It is, however, evident that this is an excellent tool to enable research in the Classics, as well as to support beginning and intermediary students of the Classics. It is also evident that the user has to apply their mind to make a selection from the list of morphological parsings, dictionary entries and word senses.

5.2 Linking in e-texts on a Kindle e-reader

The Kindle e-reader allows the reader to link to user-specified dictionaries. A large number of dictionaries is available free of charge. These dictionaries need to be downloaded by the user and the selected dictionary needs to be specified by the user when first being consulted. The specified dictionary can be changed at any time, for example, to link to either a UK or US English dictionary, or a translation dictionary, for example from English to German, or vice versa (if a German phrase occurs in the English text, or when a German text is being read). Linking occurs to the first occurrence of a lemma in the dictionary. This is usually fairly reliable, but, since there is no PoS/syntactic analysis, the landing place of the linking process is not always correct, both at the level of the morphological form (for example, verb instead of noun), the wrong lemma (in the case of homographs and homonyms), and no clarification about the correct sense of a polysemous word.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1697 (Article)

This process has been studied in a fair amount of detail in Bothma and Prinsloo 2013, where many examples are provided, as well as in Bothma and Gouws 2020, also with examples. Bothma and Prinsloo (2013: 169-184) provide a categorisation of problems that occur in the Kindle linking:

- Correct lemma but incorrect PoS (e.g. verb–noun–adjective–adverb confusion)
  - Incorrect linking of homographs, typically verb/noun confusion in dictionary linking
  - Inflected/conjugated form links to correct lemma but wrong PoS
  - Word is a homograph of an inflected/conjugated form of the verb/ noun
- Incorrect lemma (and often incorrect PoS as well)
  - Word links to a homograph lemma which is etymologically not related
  - Word links to the incorrect lemma based on the inflection of either the word itself or a possible inflected/conjugated homograph of the lemma

Bothma and Prinsloo (2013) further discuss issues with compounds, phrases and phrasal verbs, the treatment of proverbs, idioms and similar fixed expressions, the treatment of apostrophes, hyphens and capitalization, the occurrence of wrong or inappropriate options, and cases where no option is given. In many of these cases, the linking is incorrect, or requires further evaluation of the results by the user to find the appropriate sense for the context.

One example of each of the level one bullets earlier in the discussion is provided in Table 1, for ease of reference.

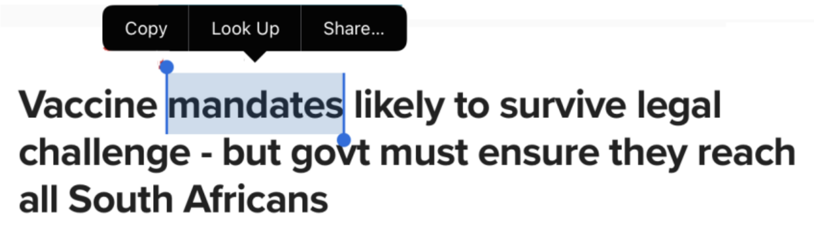
Example	Links to	Correct linking in context
"watch" as in "My watch has stopped"	**watch** ***v.* 1** [with obj.], "look at or observe attentively …"	The noun is provided later in the same article, preceded by a small black square: ▪ ***n.***
"flags", as in "He washed the flags in the courtyard"	**flag**^{1} ***n.***, "a piece of cloth or similar material …"	**flag**^{2} ***n.***, "a flat stone slab …"

Table 1: Examples of incorrect linking in Kindle

The linking of e-texts to a user-specified dictionary in the Kindle usually works very well. However, due to the fact that there is no contextual parsing, errors sometimes occur. In addition, the system is not aware of the context or cotext, which results in further possible errors. It is therefore again up to the user to apply their mind to select the correct sense or meaning in each and every instance.

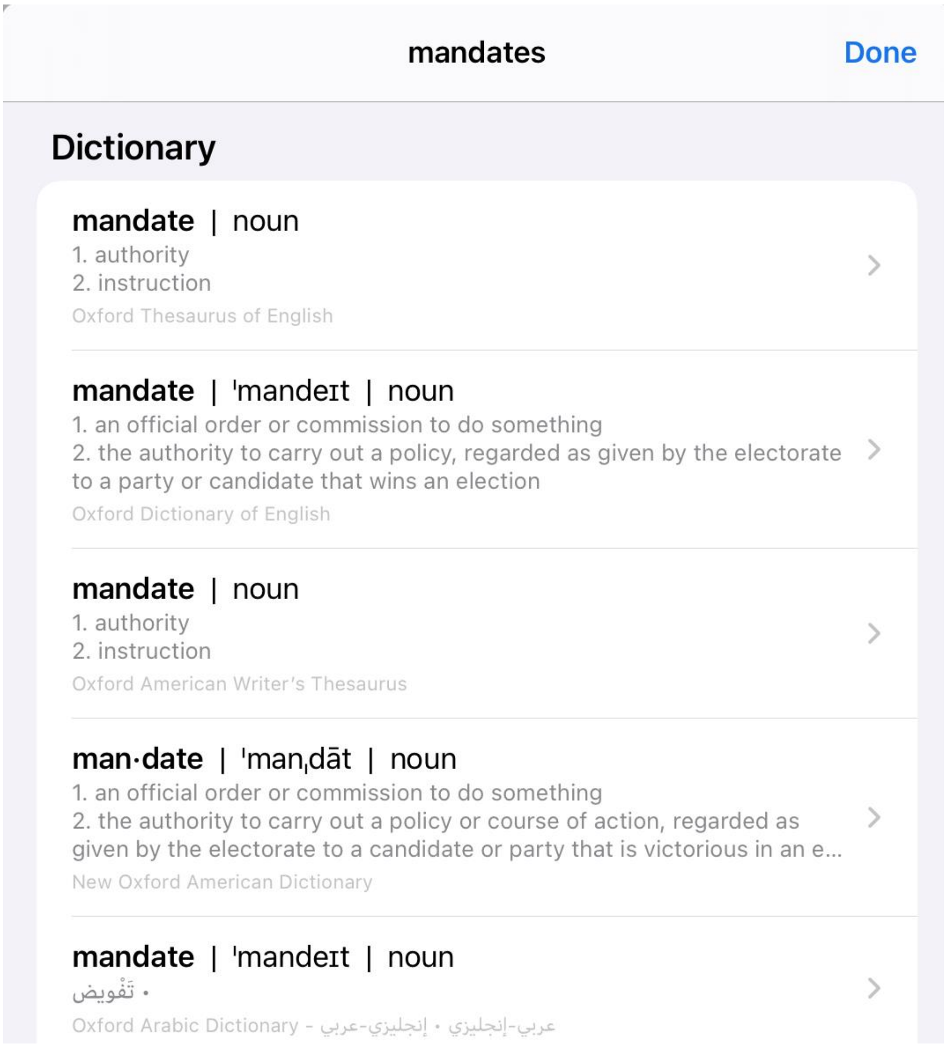
5.3 Browser-based linking

Browser-based linking from an e-text to an e-dictionary is an extension of the preceding process, in which linking is restricted to e-texts on a Kindle. In browser-based linking, any text in a browser can link to a user-selected dictionary, or set of dictionaries, which are fully customisable. This is discussed in Tarp and Gouws 2020 and in Bothma and Gouws 2020. From the examples and discussion in these two articles, it is evident that the system cannot be context-aware, and that the problems noted in the Kindle linking occur here as well. Different devices/operating systems have different interfaces, as is evident form the examples in Bothma and Gouws 2020, and Figures 10–15. Depending on the device/operating system, links to additional uncurated information sources can be accessed, which take the user outside the domain of the dictionary. One advantage of the browser-based linking on an iPad is that the user can select multiple dictionaries in the settings of their device. If a user then clicks on a word, a pop-up menu with three items is displayed (Figure 10); clicking on "Look Up" results in a customised dictionary portal with drill-down options to more information (Figure 11). This can obviously lead to information overload, but is, to a certain extent, under the control of the user, as they can select a larger or smaller number of dictionaries in the device settings. On an Android phone, however, a limited set of selection options appears (Figure 12); "Define" links to a dictionary article (Figure 13). Clicking on the three vertical dots produces a portal of unordered information sources, some of which are dictionaries, as illustrated in Figure 14. Clicking in this case on "Dictionary", a dictionary article based on the phone configuration is shown (Figure 15).



Vaccine mandates likely to survive legal challenge - but govt must ensure they reach all South Africans

Figure 10: Options when clicking on a word in Google Chrome on an iPad



mandates

Dictionary

mandate | noun
1. authority
2. instruction
Oxford Thesaurus of English

mandate | ˈmandeɪt | noun
1. an official order or commission to do something
2. the authority to carry out a policy, regarded as given by the electorate to a party or candidate that wins an election
Oxford Dictionary of English

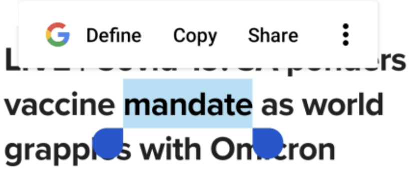
mandate | noun
1. authority
2. instruction
Oxford American Writer's Thesaurus

man·date | ˈmanˌdāt | noun
1. an official order or commission to do something
2. the authority to carry out a policy or course of action, regarded as given by the electorate to a candidate or party that is victorious in an e...
New Oxford American Dictionary

mandate | ˈmandeɪt | noun
• تَفْويض
Oxford Arabic Dictionary - عربي-إنجليزي • إنجليزي-عربي

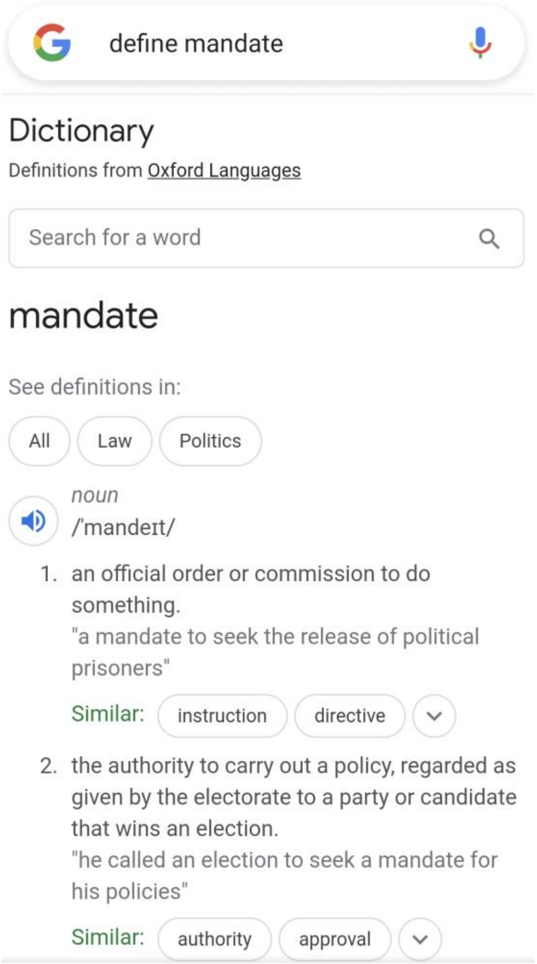
Figure 11: Dictionary portal available after clicking "Look Up" in Figure 10 in Google Chrome on an iPad

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1697 (Article)



vaccine mandate as world
grapples with Omicron

Figure 12: Options when clicking on a word in Google Chrome on an Android phone



Dictionary

Definitions from Oxford Languages

mandate

See definitions in:

noun

/'mandeɪt/

1. an official order or commission to do something.
"a mandate to seek the release of political prisoners"
Similar: instruction directive

2. the authority to carry out a policy, regarded as given by the electorate to a party or candidate that wins an election.
"he called an election to seek a mandate for his policies"
Similar: authority approval

Figure 13: Selecting "Define" in Figure 12 results in a dictionary article from *Oxford Languages*

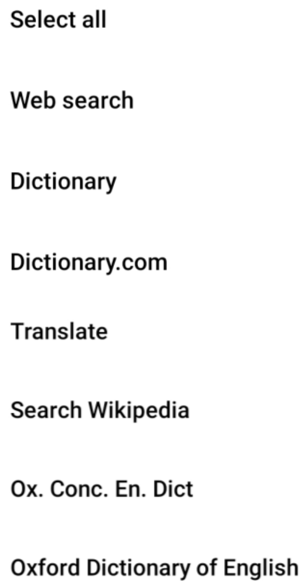
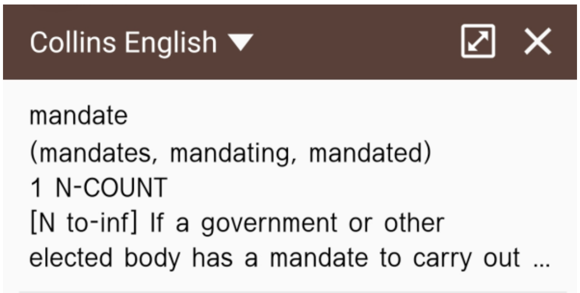


Figure 14: Clicking on the three vertical dots in Figure 12 results in this unordered portal (scrolling required to see all options)



mandate
(mandates, mandating, mandated)
1 N-COUNT
[N to-inf] If a government or other elected body has a mandate to carry out ...

Figure 15: Clicking on "Dictionary" results, in the configuration on the specific device, on a full article from *Collins English Dictionary*

In addition to the issues of potential information overload (iPad) and unordered/ illogical menu structures (Android phone), it is evident that the systems are not context sensitive and that the user has to apply their mind to obtain the correct sense or meaning in the context.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1697 (Article)

5.4 Linking to dictionaries in a language learning environment

Huang and Tarp (2021) describe *Kaiyan OpenLanguage*, an English language learning application (app), for Chinese students. The app makes use of two dictionaries, one which the authors term "embedded", and one which they term "integrated" (Huang and Tarp 2021: 74). The embedded dictionary correlates with the dictionaries described in the previous sections of this paper, and the integrated dictionary is a dictionary that is linked to the course content provided in the app, and "only the words occurring in the text can be consulted"; "Embedded dictionaries cannot 'know' what information a user is looking for in a concrete consultation, whereas integrated dictionaries, if well designed, will be context-aware and, thus, 'know' the concrete sense of a word relevant to the user". According to the authors, "This context-awareness seems to be the most urgent lexicographical challenge to the *Kaiyan OpenLanguage* and other similar learning apps" (Huang and Tarp 2021: 74-75). From their discussions in the following sections of their article, it is evident that Huang and Tarp (2021: 75-85, section 4) are not impressed with the success of context-sensitive linking in this app, and they provide a number of examples. They summarise the problems as follows:

> We have seen lexical units that are not treated in the pop-up window, and sometimes not even in the dictionaries when consulted from the front page. We have seen polysemous words where some senses are missing in the default dictionary or only available after accessing the whole article. We have seen words with inaccurate and even wrong definitions. We have seen examples of data overload with senses and equivalents that are irrelevant in the concrete context. We have seen users who have to click three or four times to get an answer or no answer at all. We have seen how the position of the pop-up window that is supposed to help users sometimes has the opposite effect and make it more difficult to pick up the meaning of a word (Huang and Tarp 2021: 86).

Contextualisation of lexicographic consultations is therefore not always very successful in this app, and Huang and Tarp (2021: 88) suggest that this should be fixed by a combination of programming and manual work, focusing on the course texts available in the app, and doing this for all texts that will be added to the app in future.

5.5 Evaluation of linking systems

Linking systems definitely simplify the dictionary consultation process. In an e-reader it is a very easy way to do a quick search in the integrated dictionary, without the "hassle" of having to go to a dictionary app or a paper dictionary.

However, currently, the quality of the results is not guaranteed because in none of the cases the system is fully aware of either the pragmatic/functional environment (the context) or the grammatical environment (the cotext). Linking systems often are not directed at the needs of a specific user group or the interpretation of a specific context. Consequently, when linked to a dictionary, the user is offered a number of options from which to choose. This often leads to a haphazard selection and the users have to apply their mind to decide what the correct meaning/sense for the given context of the word in the dictionary-external text is.

Two of the systems discussed in the article provide links between e-texts and defined corpora, viz. in the Perseus project the corpus of Latin and Greek texts (and others), and in the *Kaiyan OpenLanguage* app the course texts used for language learning.

The Perseus project has extensive markup of the texts and makes a suggestion for the correct PoS analysis based on statistical analysis, linking all possible solutions to the e-dictionaries. The user therefore still has to apply their mind to decide on the correct PoS, as well as to select the correct sense, once the dictionary is accessed.

The linking in the *Kaiyan OpenLanguage* app currently requires that the user apply their mind to make the correct selection in the pop-up window of the dictionary, very similar to the Kindle linking. The suggestions for the improvement of this system (Huang and Tarp 2021) will require extensive input from experts to ensure context-sensitive linking for the course texts.

Based on the nature of the errors in the Kindle linking, there is no analysis of the selected word, and the word is linked to the first available lemma in the dictionary The same applies to browser-based linking. There is, however, limited matching based on conjugated and inflected forms, but this is simply a matching of these items to the conjugated and inflected forms given in the dictionary, and there is no "intelligence" in the linking.

The amount of parsing in Perseus and the *Kaiyan OpenLanguage* app is insufficient to result in context-aware linking. The matching based on conjugated and inflected forms also does not lead to context-aware linking. The only currently available option is the laborious manual linking by experts. This can obviously be done only for a limited/small corpus, and other options should be investigated.

This article suggests the partial-automated and manual construction of a fully annotated (marked-up) corpus of limited size, to serve as the input for a machine-learning system with artificial intelligence. We therefore foresee that there will be a "black box" of software between the selected word and the dictionary that will determine the correct lemma and sense to be selected from the e-dictionary.

Based on the analyses of the four systems discussed thus far in this article, it is evident that, in all cases, the systems are not context aware, either in terms of the context/cotext of the item or in the broader context of the text:

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1697 (Article)

- Context/cotext issues relate to the system not being aware of, *inter alia*:
  - the PoS of an item;
  - the syntactic function of the item;
  - a distinction between homographs
  - the sense in context of a polysemous word

- Broader context issues of the text relate to, *inter alia*:
  - the location in which the text is situated
  - the time period in which the text is situated
  - the style of the text, or the specific portion of the text, e.g. formal, informal, slang

In short, it is evident that linking systems need to be devised for a better pairing with specific dictionaries and dictionary entries, based on the context and cotext of the word in the text.

In the rest of this article, we differentiate between linking in a corpus with markup and linking in free text. Modular components and characteristics of linking software will also be discussed.

6. The construction of a corpus of limited size

When devising the linking system that helps a user to move from a word in a dictionary-external text to the appropriate item in the dictionary, it is necessary to start with the markup of a small corpus to enable the initial machine learning processes on the side of the e-device and its software. All texts to be downloaded onto the e-device cannot be marked-up and the ideal situation proposed in this paper is that the software of the e-reader or the browser will eventually be able to identify the context of a word in such a way that the linking to the dictionary can be done in an unambiguous way.

When one works with a small, defined corpus, it is possible to annotate the documents with metadata. Fine-grained metadata are required to describe the text sufficiently. This includes:

- PoS tagging, syntactic dependencies and certain semantic aspects;
- bibliographic detail, especially indicating the full volume as well as in-text occurrences
- functional/pragmatic data, including an indication of style, register and language varieties

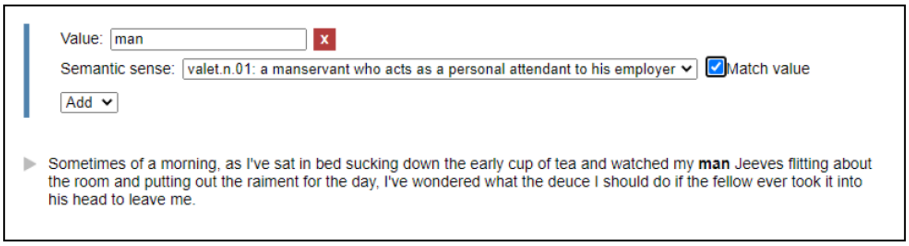
The linking process should then match the metadata of the word in a text with a dictionary article, and with a specific item or search zone in the dictionary article. The markup we propose to achieve this, is similar to markup for enhanced retrieval of specific words and phrases from a text, as described in

Ball (2020: 160-188). In this case, a specialised search engine retrieves a word or phrase that has specific attributes, by matching these attributes to the markup of the words in the database. Ball (2020) developed a prototype system to illustrate these functionalities; for details, see Ball (2020: 198-242).

The matching of a search string with specialised metadata markup to a word or phrase with identical markup in the text database (marked-up corpus) is similar to matching a word with specialised metadata markup in an e-text to items with identical markup in the dictionary database. The database structures will obviously have to be adapted, but the underlying principles are similar. The examples based on retrieving items from a corpus are therefore relevant to illustrate the principle of matching items in an e-text to items based on identical metadata in a dictionary database.

Two examples are selected from Ball (2020) to illustrate such matching, in these cases searching according to a specified semantic sense, and searching according to functional properties, viz. the language of the word.

Figure 16 illustrates a search for the value "man" with the sense of "a man servant who acts as a personal assistant to his employer" (Ball 2020: 219). All other cases of "man" in the texts were not retrieved, and only the one with the required sense was retrieved. This sense was manually marked up in the database.



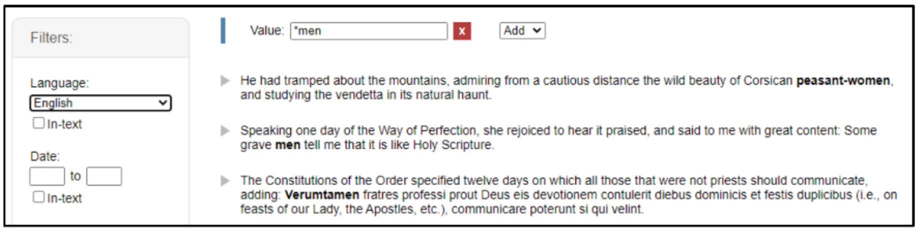
Value: man

Semantic sense: valet.n.01: a manservant who acts as a personal attendant to his employer ☑ Match value

Sometimes of a morning, as I've sat in bed sucking down the early cup of tea and watched my **man** Jeeves flitting about the room and putting out the raiment for the day, I've wondered what the deuce I should do if the fellow ever took it into his head to leave me.

Figure 16: Searching for the value "man" with the sense of "a man servant who acts as a personal assistant to his employer" (Ball 2020: 219)

In the next three figures. the search was for the truncated form "*men". In Figure 17, the language of the volume was specified as English, and three cases were retrieved, two cases of English words, and one of a Latin word; in Figure 18, English was specified as the "in-text" language, and only the English examples were retrieved; in Figure 19, Latin was specified as the in-text language, and only the Latin example was retrieved (Ball 2020: 230).


Filters:

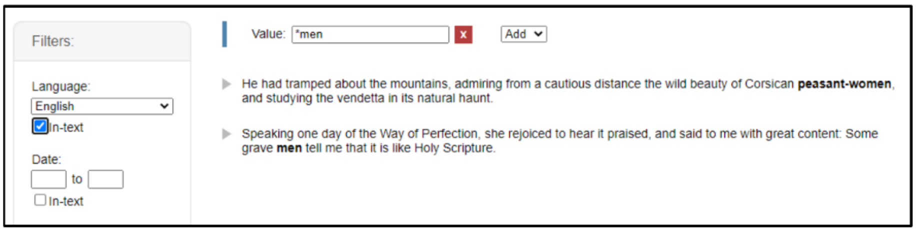
Language: English
☐ In-text

Date: to
☐ In-text

Value: *men

- He had tramped about the mountains, admiring from a cautious distance the wild beauty of Corsican **peasant-women**, and studying the vendetta in its natural haunt.
- Speaking one day of the Way of Perfection, she rejoiced to hear it praised, and said to me with great content: Some grave **men** tell me that it is like Holy Scripture.
- The Constitutions of the Order specified twelve days on which all those that were not priests should communicate, adding: **Verumtamen** fratres professi prout Deus eis devotionem contulerit diebus dominicis et festis duplicibus (i.e., on feasts of our Lady, the Apostles, etc.), communicare poterunt si qui velint.

Figure 17: Searching for the truncated form "*men", with English as the language of the volume (Ball 2020: 230)


Filters:

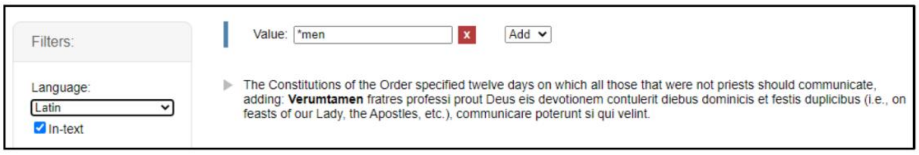
Language: English
☑ In-text

Date: to
☐ In-text

Value: *men

- He had tramped about the mountains, admiring from a cautious distance the wild beauty of Corsican **peasant-women**, and studying the vendetta in its natural haunt.
- Speaking one day of the Way of Perfection, she rejoiced to hear it praised, and said to me with great content: Some grave **men** tell me that it is like Holy Scripture.

Figure 18: Searching for the truncated form "*men", with English specified as the in-text language (Ball 2020: 230)


Filters:

Language: Latin
☑ In-text

Value: *men

- The Constitutions of the Order specified twelve days on which all those that were not priests should communicate, adding: **Verumtamen** fratres professi prout Deus eis devotionem contulerit diebus dominicis et festis duplicibus (i.e., on feasts of our Lady, the Apostles, etc.), communicare poterunt si qui velint.

Figure 19: Searching for the truncated form "*men", with Latin specified as the in-text language (Ball 2020: 230)

It is important to assess the role of a granular metadata markup of textual data in establishing context, as illustrated in the previous examples. Two significant aspects in this regard are the volume context and the section context. Volume context refers to bibliographic metadata, including a reference to the author and other relevant data. This type of context also includes functional/pragmatic metadata, for example style, location, dialectal, cultural data. On the other hand, section context refers to grammatical metadata, in-text biblio-

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1697 (Article)

graphic metadata and functional/pragmatic metadata, for example, direct speech, style, location, dialectal, cultural data, and could also include functional/pragmatic metadata applicable to the paragraph or sentence.

This process can be partially automated, specifically for PoS tagging, and to a lesser extent for syntactic tagging, but all markup needs to be checked manually to ensure that the markup data are correct. Tagging at the semantic, functional and bibliographic levels cannot yet be semi-automated (Ball 2020: 194).

A relevant question is to what extent can such and more complex mappings be automated, and what technologies are required to do so? According to Tarp and Gouws (2019: 264) "… the long road to the required data means that full contextualization is still a challenge to modern lexicography …" They state that comprehensive research is done "in order to develop a tool that can deduce the specific meaning of a word from the context."

Successful mapping will require a variety of parsing technologies and some of these applications will depend on the progress already made with regard to the specific language. For English a lot of success has been achieved but there are some unsolved challenges. The situation regarding English is currently as follows:

- Automated PoS tagging is good
- Automated syntactic dependencies are in general fair
- Automated semantic tagging is poor
- Automated bibliographic tagging is not yet possible
- Automated functional, pragmatic tagging is not yet possible.

It is important to note that tagging is not an isolated process, but certain interrelationships are important. Semantic tagging depends on accurate context, syntactic, functional tagging and selection restrictions, whereas syntactic tagging depends on accurate PoS and valency tagging. In addition, accurate taggers at all levels are essential if one wants to automate tagging.

Full automation is therefore currently not yet possible, as described in Ball (2020: 243-279). Semi-automated and manual coding are very laborious and time-consuming tasks, and with current technologies it is not possible to encode very large data sets accurately. We therefore suggest that a fairly small corpus be annotated in detail, using currently available tools, and checking them very carefully manually, to ensure data quality and data integrity.

Based on the preceding descriptions, a limited corpus of annotated text can be created. On-the-fly tagging and analyses cannot simulate the functionalities of such a corpus, but the detail markup in the corpus is required for contextualization. Based on such a corpus, machine learning and artificial intelligence software can be trained to deduce many of these features for proper contextualization automatically. Such software, in a "black box" between the e-text and the e-dictionary can therefore facilitate proper contextualization. A number of further issues are addressed in the following sections of this paper.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1697 (Article)

7. Does a text need to have formal indications about context?

The success of mapping procedures between an e-reader (or similar device) and an e-dictionary demands innovative adaptations in both the e-reader and the e-dictionary that will result in a new perspective on the procedure of contextualization.

In the lexicographic treatment of words taken from texts, the items in the dictionary articles are complemented by items giving contextual and cotextual guidance. These complementing items may not be chosen at random but should be taken from corpus-based data that reflect the actual use of the linguistic expressions. For successful mapping between a dictionary-external text and the appropriate subcomment on semantics in a dictionary article, the notion of parallel contexts and parallel contextualization needs to be negotiated. This does not imply a marking-up of every text downloaded onto the e-device, although when only dealing with a small corpus formal indications of contexts are feasible and would be required to ensure accurate mappings of meanings and senses. Where such a corpus is compiled for language for general purposes, context will be difficult to deduce, as there are no formal markers. In the case of languages for special purposes it might be easier because the topic or discipline might offer limited context.

However, this cannot be done for each text read on an e-device. Parallel contextualization does not in the first instance prevail between the integrated dictionary and each individual text on the e-device, but between the dictionary and the software of the e-reader. Consequently, successful mapping does demand much stronger contextual considerations in the software of the e-reader or other device. This enhanced context awareness of the e-reader or other device could enable a better linking between a word in a text on that device and not only a lemma sign in a dictionary article that represents that specific word, but the specific item in such a dictionary article that presents the applicable treatment of the word in the dictionary-external text.

Parallel contextualization should be preceded by another phase in the integration of a dictionary into a device like an e-reader, namely a determining of the items giving context and cotext in the dictionary. The macrostructural coverage of the dictionary should be corpus-based and must reflect the active lexicon of the treated language as well as some items with a lesser usage frequency. An important early phase in the movement towards parallel contextualization is to ensure that the treatment in the dictionary displays a thorough account of the typical cotexts and contexts in which the treated words occur. This emphasises the significance of the correct choice of dictionaries to be integrated. The user-perspective, a dominant criterion in modern-day lexicography, should also be transferred to the e-device. The users will be primary users of the e-device and only secondary users of the integrated dictionary. Therefore, the candidate dictionary should not be randomly chosen only on account

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1697 (Article)

of its lexicographic quality, important as it will always be, but the selection should be motivated by the intended use and users of the e-device.

Typical texts to be read on the e-reader (or similar device) will play a determining role in the selection of the dictionary that has to be integrated. Once this has been determined a context parallel to that of the dictionary needs to be accounted for in the software of the e-reader. The e-reader should contain context guidance that parallels that of the dictionary. When a reader clicks on a given word in a text on the e-reader the software interprets the context of that word and matches it with the contexts found in the dictionary article that has that word as lemma sign. This could then lead to successful mapping.

Each text encountered on the e-device does not need to have formal indications about context. Such indications should be found in the e-device. This mapping would still be difficult when dealing with free texts. The analysis of non-stop words in the text could probably give an indication, but, ever so important, the possible role of AI and machine learning in ensuring successful mapping should be investigated.

8. Should the cotext be negotiated?

Comparable arguments given for context could also be given for cotext. The software of the e-device should be able to analyse a limited section to be able to identify the relevant cotext of the section. In a corpus it has to be based on the markup of paragraphs or sentences, for example, in-text citation, direct speech, collocations and other metadata. In free text, probably the same features will apply, if sufficient metadata can be deduced by the software. The cotext in which a word occurs is important to ensure the correct mapping because the occurrence of a word in a text participates in activating a specific homonym/homograph or a specific sense of a polysemous word. The correct analysis of the cotext in the text could ensure the correct mapping with a specific item in the dictionary. The accurate linking of the cotext can enhance the speedy retrieval of information from the data presented in the dictionary article.

9. The structure of the dictionary database

Integrating a dictionary and an e-device implies that the dictionary becomes an instrument that must satisfy the lexicographic needs of the users of the e-device. The software linking the e-device to the e-dictionary will need to be adapted to enable the coordination of cotexts and contexts. Possible changes, including changes to the database of the dictionary, also need to be negotiated at an early phase of the marriage between dictionary and e-device.

Once integrated into the e-device the dictionary loses its independent status and use. Consultation is no longer open because the integrated dictionary has restricted access possibilities and can only be accessed via a click on a

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1697 (Article)

word in a text on the e-device. If an existing dictionary is linked to the e-device the complexity of its former database will determine whether the database has to be changed. If the database contains a sufficient number of unique record types and a sufficient number of properties/attributes, it would most probably not be necessary to change the database structure. However, if there are insufficient record types and attributes, the structure should be adapted to make provision for these items. Nevertheless, one should be careful of over-complicating the database.

10. What are modular components and characteristics of such software?

Negotiating improved contextualization possibilities compels an investigation of the modular components and characteristics of the envisaged software. The dictionary and the e-device function as a unit and we foresee that a black box of software will be running in the background, to do the required analyses and mappings. The software in the black box will include PoS and other taggers, NLP technologies as well as AI and machine learning.

The software will interpret the text and interact with the dictionary database by mapping the attributes of the word in the text with the attributes of different fields in the database. If the mapping is successful, the user will be provided with a contextualized result.

However, it is important to note that the software should require no or limited manipulation by the user.

11. Scaling the data set

When embarking on a free text implementation it is a prerequisite to scale the data set and to have a gradual increase in the application possibilities. Starting with a small data set for markup, the markup can be partially automated. Fully automating the markup is problematical, because of inaccuracies at various levels, as discussed in previous sections. It will be essential to check the accuracy of all markup procedures manually to ensure quality. Working with only a small data set is very restrictive and any serious research in this environment should move from a defined corpus to general e-texts, to improve the quality and efficiency of the machine learning algorithms.

12. Future research

Most of what we have stated in this article requires further research. Some questions that could guide such research include whether contextual analysis should be provided for every word in a text? Unless the software of the e-device can link the context of the text to the appropriate search zone in a specific dictionary article, the answer would probably be yes. A follow-up ques-

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1697 (Article)

tion is whether this would be based on a full set of parsing? This will most probably be required for an efficient disambiguation of homographs, etc. Further research could also result in presenting drill down options that could allow the user to retrieve other information types, like items giving morphology, inflection, collocations, etc.

Such an amount of work raises the question of processing overload. To a certain extent this depends on implementation efficiency, but also on device capabilities. Processing should be fast enough that the user is not frustrated by the lack of speed of the system. In terms of what current technologies offer, better quality parsers, specifically for syntactic, semantic and functional issues, are required. In addition, the technical specifications of artificial intelligence and machine learning algorithms need to be addressed, taking cognizance of the technological challenges and also the financial implications. Attempts should also be made to distinguish between what is traditionally known as a dictionary and what could in future perhaps be referred to as a lexicographical database.

11. In conclusion

Contextualization and cotextualization are essential goals that need to be addressed in future research. Extensive multi-disciplinary research and collaboration are required. Lexicographers, computer and information scientists, NLP and UX specialists and others will have to collaborate. It is going to be complex, and an easy solution is not envisaged. When attempting to develop genuine smart e-dictionaries, we concur with the final comment of the article by Tarp and Gouws (2019: 266): "We have work to do."

Bibliography

Ball, L.H. 2020. *Enhancing Digital Text Collections with Detailed Metadata to Improve Retrieval.* PhD dissertation, University of Pretoria. Available: https://repository.up.ac.za/handle/2263/79015 (accessed 18 June 2021).

Bothma, T.J.D. and R.H. Gouws. 2020. e-Dictionaries in a Network of Information Tools in the e-Environment. *Lexikos* 30: 29-56.

Bothma, T.J.D. and D.J. Prinsloo. 2013. Automated Dictionary Consultation for Text Reception: A Critical Evaluation of Lexicographic Guidance in Linked Kindle e-Dictionaries. *Lexicographica* 29(1): 165-198.

Crane, G. 1998. The Perseus Project and Beyond: How Building a Digital Library Challenges the Humanities and Technology. *D-Lib Magazine* 4(1). Available: https://www.dlib.org/dlib/january98/01crane.html (accessed 18 June 2021).

Domínguez, M.J. and R.H. Gouws. (In print). Regarding the Definition, Presentation and Automatic Generation of Contextual Data in Lexicography.

Engelberg, S. and C. Müller-Spitzer. 2013. Dictionary Portals. Gouws, R.H. et al. (Eds.). 2013. *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography.* Berlin: De Gruyter: 1023-1035.

Gouws, R.H. 2002. Equivalent Relations, Context and Cotext in Bilingual Dictionaries. *Hermes — Journal of Language and Communication in Business* 15(28): 195-209.

Gouws, R.H. 2021. Expanding the Use of Corpora in the Lexicographic Process of Online Dictionaries. Piosik, M. et al. (Eds.). 2021. *Korpora in der Lexikographie und Phraseologie:* 1-20. Berlin: De Gruyter.

Gouws, R.H. and Prinsloo, D.J. 2005. *Principles and Practice of South African Lexicography.* Stellenbosch: African SUNMedia.

Huang, F. and S. Tarp. 2021. Dictionaries Integrated into English Learning Apps: Critical Comments and Suggestions for Improvement. *Lexikos* 31: 68-92.

Lettner, K. 2020. *Zur Theorie des lexikographischen Beispiels.* Berlin/Boston: De Gruyter.

Perseus Digital Library. n.d.-a. *Greek and Roman Documents*. Available: http://www.perseus.tufts.edu/hopper/collection?collection=Perseus:collection:Greco-Roman (accessed 12 October 2021).

Perseus Digital Library. n.d.-b. *Perseus Digital Library — About*. Available: http://www.perseus.tufts.edu/hopper/about (accessed 12 October 2021).

Perseus Digital Library. n.d.-c. *Perseus Digital Library — Browse the Collections*. Available: http://www.perseus.tufts.edu/hopper/collections (accessed 12 October 2021).

Rydberg-Cox, J.A., R.F. Chavez, D.A. Smith, A. Mahoney and G.R. Crane. 2000. Knowledge Management in the Perseus Digital Library. *Ariadne* 25. Available: http://www.ariadne.ac.uk/issue/25/rydberg-cox/ (accessed 18 June 2021).

Tarp, S. 2012. Theoretical Challenges in the Transition from Lexicographical p-works to e-tools. Granger, S. and M. Paquot (Eds.). 2012. *Electronic Lexicography*: 107-118. Oxford: Oxford University Press.

Tarp, S. and R.H. Gouws. 2019. Lexicographical Contextualization and Personalization: A New Perspective. *Lexikos* 29: 250-268.

Tarp, S. and R.H. Gouws. 2020. Reference Skills or Human-Centered Design: Towards a New Lexicographical Culture. *Lexikos* 30: 470-498.

Wiegand, H.E. 1988. Wörterbuchartikel als Text. Das Wörterbuch. Artikel und Verweisstrukturen. G. Harras (Ed.). 1988. *Jahrbuch des Instituts für deutsche Sprache* 1987: 30-120. Düsseldorf: Schwann.

Wiegand, H.E. 1998. *Wörterbuchforschung.* Berlin: De Gruyter.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1698 (Article)

Applied Corpus Linguistics for Lexicography: Sepedi Negation as a Case in Point

Gertrud Faaß, *Department of Information Science and Natural Language Processing, University of Hildesheim, Germany (gertrud.faass@uni-hildesheim.de)*

Abstract: So far, Sepedi negations have been considered more from the point of view of lexicographical treatment. Theoretical works on Sepedi have been used for this purpose, setting as an objective a neat description of these negations in a (paper) dictionary. This paper is from a different perspective: instead of theoretical works, corpus linguistic methods are used: (1) a Sepedi corpus is examined on the basis of existing descriptions of the occurrences of a relevant verb, looking at its negated forms from a purely prescriptive point of view; (2) a "corpus-driven" strategy is employed, looking only for sequences of negation particles (or morphemes) in order to list occurring constructions, without taking into account the verbs occurring in them, apart from their endings. The approach in (2) is only intended to show a possible methodology to extend existing theories on occurring negations. We would also like to try to help lexicographers to establish a frequency-based order of entries of possible negation forms in their dictionaries by showing them the number of respective occurrences. As with all corpus linguistic work, however, we must regard corpus evidence not as representative, but as tendencies of language use that can be detected and described. This is especially true for Sepedi, for which only few and small corpora exist. This paper also describes the resources and tools used to create the necessary corpus and also how it was annotated with part of speech and lemmas. Exploring the quality of available Sepedi part-of-speech taggers concerning verbs, negation morphemes and subject concords may be a positive side result.

Keywords: AFRICAN LANGUAGES DICTIONARIES, CORPUS LINGUISTICS, NEGATION, SEPEDI, NORTHERN SOTHO, LEXICOGRAPHY, PART-OF-SPEECH TAGGING, CORPUS QUERY PROCESSING

Zusammenfassung: Eine korpuslinguistische Untersuchung der Sepedi-Negation für die Lexikographie. Bisher wurden Sepedi Negationen eher aus der Sicht der lexikographischen Behandlung betrachtet. Hierfür wurden theoretische Werke über Sepedi verwendet, wobei als Zielsetzung eine saubere Beschreibung dieser Negationen in einem (Papier-)Wörterbuch gesetzt wurde. Dieser Beitrag ist aus einer anderen Perspektive: statt theoretischer Werke werden korpuslinguistische Methoden eingesetzt: (1) ein Sepedi Korpus wird auf Basis bestehender Beschreibungen zu den Vorkommen eines einschlägigen Verbs untersucht und dabei seine negierten Formen aus rein präskriptiver Sicht betrachtet; (2) wird eine "corpus-driven"-Strategie eingesetzt, bei dem nur nach Sequenzen von Negationspartikeln (oder Morphemen) gesucht wird, um vorkommende Konstruktionen auflisten zu können, ohne dabei die dabei vorkommenden Verben —

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1698 (Article)

abgesehen von ihrer Endung — zu berücksichtigen. Der Ansatz in (2) soll dabei nur eine mögliche Methodik aufzeigen, um bestehende Theorien über vorkommende Negationen erweitern zu können. Wir möchten auch versuchen, Lexikographen darin zu unterstützen, eine frequenzbasierte Reihenfolge der Einträge möglicher Negationsformen in ihren Wörterbüchern aufzustellen, in dem wir ihnen die Anzahl der jeweiligen Okkurrenzen aufzeigen. Wie bei allen korpuslinguistischen Arbeiten müssen wir jedoch Korpusevidenz nicht als repräsentativ ansehen, sondern als Tendenzen des Sprachgebrauchs, die festgestellt und beschrieben werden können. Dies gilt insbesondere für Sepedi, für das nur wenige und kleine Korpora existieren. Dieser Beitrag beschreibt außerdem die Ressourcen und Werkzeuge, die verwendet wurden, um das nötige Korpus zu erstellen und auch, wie dieses mit Wortart und Grundformen der Wörter angereichert wurde. Ein Nebenergebnis ist dabei die Untersuchung der Qualität von verfügbaren Taggern bzgl. Verben, Negationsmorphemen und Kongruenzpartikel

Stichwörter: WÖRTERBÜCHER AFRIKANISCHER SPRACHEN, KORPUSLINGUISTIK, NEGATIONEN, SEPEDI, NORD-SOTHO, LEXIKOGRAPHIE, TAGGING, BEARBEITUNG VON KORPUSANFRAGEN

1. Introduction

Negation is an important issue in language description because of the multiple forms in which it may occur. Attempts have been made to categorize negation alongside traditional linguistic fields like morphology (word level) and syntax (phrase level). Dahl (1979: 81), for example, initially distinguishes morphological and syntactical negation, describing syntactic negation as using "simple and double particles, negative auxiliaries and particle + dummy auxiliaries" while he sees negation on a morphological level as part of inflection. Yet he later adds (ibid: 83) that "in most cases" there is no such clear-cut description possible and that the distinction can only be made between negation morphemes as affixes (bound morphemes) or as particles (free morphemes).

This contribution investigates negation in the language Northern Sotho (ISO-code: NSO), also called Sepedi. The work is based on Prinsloo (2020), who describes the "lexicographic treatment of Sepedi negations" from the view of lexicography. In his article, he lists a number of forms that negation takes on, using theoretical work on Sepedi as knowledge base with the aim of properly describing Sepedi negation in a paper dictionary. Here, we will attempt to explore corpus data in a semi-automated way, utilizing existing Natural Language Processing (NLP) tools along the way.

There are two approaches to examine corpora: *corpus-based* and *corpus-driven* (Tognini-Bonelli 2001). The former starts with theoretical hypotheses about a language and investigates whether these are true, while the latter explores phenomena that are significant from a quantitative point of view (e.g. word sequences appearing frequently) to find new insights into the use of a language. In this paper, we start with a corpus-based approach by querying a cor-

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1698 (Article)

pus of pre-defined negated verb formations. Interestingly enough, we also come across new formations not defined in the literature.

NLP processing usually begins with tokens matching language units. As most NLP tools were initially developed for European languages, it is usually assumed that one token is either to be identified as a symbol (like punctuation) or equivalent to one word (which means: a free morpheme). For the South African indigenous languages, this assumption is however often not true. Sepedi, for example, utilizes the so-called "disjunctive writing system", that is bound morphemes are often written separately from the morpheme they belong to, hence negation affixes (bound morphemes) and negation particles (free morphemes) cannot be distinguished when tokenising. The tools therefore treat both as if they were particles.

Our approach is to work with language in use, so it is necessary to compile a corpus containing as many freely available sentences as possible. Before this corpus can be investigated properly, however, its tokens should be annotated with their respective part-of-speech (POS), so that queries can be performed on that level (e.g. to find all verbs). An annotation of lemmas (in the sense of a base form of each word) might also prove helpful, especially for languages with a rich morphology.

For the purpose of sentence collection, we make use of data available from the South African Centre for Digital Language Resources (SADiLaR[1]) and the Sepedi corpora collected by the CURL web collection machine located at Leipzig University (Goldhahn et al. 2016). For corpus annotation, NLP tools provided by SADiLaR and an own Sepedi tagging parameter file (Faaß et al. 2009), developed for the TreeTagger (Schmid 1994; Schmid 1995) are utilized. The corpus is encoded within the freely available IMS Open Corpus Query Workbench (OCWB) system (Evert and Hardie 2011), and the respective queries for the corpus are written as macros for the two purposes of a better documentation and reproducibility.

2. Aims

Corpus linguistics is not just a science in its own right; it can also be seen by other research fields as a helpful method that can be applied for their empiric research. In lexicography, utilizing corpus data is essential (see e.g. Faaß 2018).

In corpus linguistics, the language in use needs to be described and the resources utilized should contain utterances by as many speakers as possible, to at least get a grip on how certain linguistic phenomena appear in the living language. Frequencies of occurrences found in corpora assist in deciding which linguistic phenomena and/or word forms should preferably be included in a dictionary because a user of a dictionary — often a learner of the language — should find at least the most frequent word forms.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1698 (Article)

Concerning Sepedi, there are often several ways to negate a verb, therefore the first aim of this contribution is finding the most frequent negation strategies of a selected verb[2] as a case in point with the aim of showing them in the right order in a dictionary entry. Secondly, Prinsloo (2020) describes the frequencies of occurrences of single negation morphemes; this paper will add morpheme (or particle) sequences appearing in the corpus thereby widening the issue to a morpho-syntactical description of negation strategies utilized. Although Prinsloo (ibid.) also dedicates a chapter to copulatives, this contribution focuses on full verbs.

3. Resources and their utilization

Starting from collecting a corpus of Sepedi sentences, we proceed with tokenising the texts and detecting sentence borders. Subsequently, the tokens are labelled (annotated) with POS utilizing two freely available taggers. A Sepedi lemmatiser is then used to add lemmas as additional labels to the tokens. The resulting corpus is encoded in the IMS Open CorpusWorkBench (OCWB, version 3.4.32) to ease its exploration. Lastly, the corpus is queried with the help of written macros to ensure reproducibility of results.

3.1 The SEPEDI2021 Corpus

The first task when collecting a corpus is to find as many utterances of the language as possible. Often, such resources are made available by repositories. The virtual language observatory hosted by the Common Language Resources and Technology Infrastructure, CLARIN[3] shows that there are data available from SADiLaR[4] and from the University of Leipzig. Additionally, Leipzig offers CURL (Goldhahn et al. 2016), an online web crawler tool into which URLs of web pages containing Sepedi text can be fed. After the fully automated crawling and pre-processing is completed, the resulting corpus is made available for download. CURL was already processed in 2017 generating a small Sepedi corpus, and again, with newer URLs collected by the author of this contribution in 2021.

We built our corpus using these available resources: SADiLaR's Sepedi Text corpus forms the biggest part of our corpus (Eiselen and Puttkammer 2014). In the Sepedi Speech corpus (De Vries et al. 2014), we change the boundary mark <orth> to <s>, that is count utterances as sentences in line with the other parts of our corpus. Adding the two CURL corpora, we compile a corpus of 154,204 sentences (2.7 million tokens), as shown in Table 1.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1698 (Article)

Name of the resource	*repository*	*no. of unique …*	*no. of tokens*
NCHLT Speech Corpus[5]	SADiLaR	56,284 utterances	238,905
NCHLT Text Corpus[6]	SADiLaR	83,614 sentences	2,224,593
NSO_Community 2017	Leipzig Wortschatz Project	4,746 sentences	113,392
NSO-Community 2021	Leipzig Wortschatz Project	9,560 sentences	178,005
Total		154,204 sentences	2,754,895
SEPEDI2021		81,274 sentences	2,327,390

Table 1: Parts of the SEPEDI2021 corpus

However, when using textual data from different sources, doublets must be expected. Therefore, we sort all sentences uniquely before further processing. We also manually delete sentences which have been collected even though they contain several words from other languages. The resulting SEPEDI-2021 corpus consists of 69,439 unique sentences (near doublets were not deleted). Lastly, we run a local tokeniser (its output is a one token per line format) that adds a number of sentence borders (some lines with "sentences" of the provided text corpora contained several sentences). We also change all occurrences of more than three dots into "…" (that is one token). Counting the output, we find that our final SEPEDI2021 corpus in total contains 2,327,390 tokens in 81,274 sentences.

3.2 Tagging the corpus with parts-of-speech (POS)

Unfortunately, the POS-tagger parameter file (Taljard et al. 2008, and Faaß et al. 2009) produced in 2009 for the rft-tagger (Schmid and Laws 2008) can only be used on 32-bit machines which have been replaced during the past decade with 64-bit machines. Using that parameter file, the RFT-Tagger achieved 94.16% precision (Faaß et al. 2009). The training material is no longer available, we therefore have to make use of the alternative and still usable TreeTagger parameter file reaching 92.46% precision (ibid.), using the label "tpos". The SADiLaR (NCHLT) tagger by Eiselen and Puttkammer (2014), claims a tagging precision of 96%, therefore we annotate its annotation as "npos" to the corpus, as well[7].

3.3 Lemmatising the corpus

Eiselen and Puttkammer (2014) also provide an NCHLT lemmatiser[8], a tool

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1698 (Article)

generating base forms for inflected word forms. However, applying this tool is different from similar ones: instead of directly utilizing it on a running text, one must provide the tool with a lowercase word list containing the words of the corpus (to save execution time, the list should be sorted uniquely beforehand). This way of processing is unusual — lemmatising like tagging is usually a process of looking at words in their context (seeing that many words are ambiguous and thus may have several lemmata to choose from). To annotate the lemmata which were identified in the corpus, it is necessary to write a tool using the output of the lemmatiser as an inventory. Examining the results of the lemmatiser, we note that many inflected words (especially verbs) of the corpus' word list were not lemmatised but remained in their original form. It seems that wherever there a lemma is unknown, the tool uses the word itself. Before adding these word forms, we change all characters to lower case to have at least an entry in the lemma field and to facilitate the formulation of queries at a later stage.

3.4 Corpus Annotation Overview/Encoding

The resulting corpus, ready to be encoded with the IMS Open CorpusWork-Bench (Evert and Hardie 2011), has a table format, containing the columns "word" for the original token, "tpos" (for tree-tagger POS), "npos" (for NCHLT POS and lemma (which might be the lower case version of the token). Note that the column titles of Table 2 will be utilized as attributes in the queries described from section 4 on.

Table 2 shows a small excerpt of the corpus, demonstrating the ambiguity of *a*, which according to the taggers in its first appearance is either a particle or a subject concord of noun class 1 (the npos annotation is correct). Both taggers annotate it as morpheme (of the present tense) in its second appearance (a table listing the tags utilized by both taggers can be found in the Appendix).

word	*tpos*	*npos*	*lemma*
<s>			
A	PART	CS01	a
pudi	N.09	V	pudi
re	CS.PERS	CSPERS	re
a	MORPH	MORPHPRES	a
feditše	V	V	feditše
.	$.	ZE	.
</s>			

Table 2: An example SEPEDI2021 sentence ready for encoding

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1698 (Article)

4. Utilizing the Corpus Query Processor

4.1 General Information

With the IMS Open CorpusWorkBench (Evert and Hardie 2011), the tool Corpus Query Processor (CQP) is provided. This tool can easily be used to query corpus data not only on word level, but also on each of the annotation levels provided by the encoder by way of attribute-value constraints. The queries work on corpus positions (each containing one token and its annotations). A query for such a position is bound by "[]" (if not filled, all tokens in the corpus are found with this query). The query [word="a"] finds all occurrences of the token *a*, while a combination with the npos-attribute [(lemma="a") & (npos="CS01")] finds all upper and lower case occurrences of *a* being annotated by the NCHLT tagger as a subject concord of class 1.

In our corpus, the general query searching for the lemma *a* without any further constraint finds 67,623 occurrences. *a* is a highly ambiguous morpheme (see also Faaß et al. 2009). Table 3 shows the annotations and their frequencies as identified by the two taggers in the SEPEDI2021 corpus.

Annotation *NCHLT / TreeTagger*	*NCHLT tagger* *("npos")*	*TreeTagger* *("tpos")*
CPOSS06 / CPOSS.06	22,583	18,046
CS01 / CS.01	21,472	32,292
CS06 / CS.06	15,466	8,130
CD06 / CDEM06	4,335	3,605
MORPHPRES / MORPH	2,350	2,927
TENSE / MORPH	29	n.a.
PART / PART	451	1,972
CO06 / CO.06	440	628
RV (wrong tag)	152	—
CD01 / CDEM.01	146	0
RS (wrong tag)	112	—
VCOP / VCOP	54	5
QUE / QUE	33	0
Total	67,623	67,623

Table 3: Occurrences of *a* in the Sepedi2021 corpus with its npos- and tpos-annotations

CQP also allows for querying sequences of tokens, we can thus for example query the sequences of negation morphemes again on all available levels of annotation. Marking structural annotations like sentence borders in our queries ensures that we remain within a sentence when querying sequences. As an important advantage of this tool, we may make use of regular expressions (RegEx) when describing values to match. Using RegEx shortens the necessary processing time significantly. Therefore, we may describe the set of regular subject concords *ke, re, le, se, e, bo, go, o, ba, a, di,* in the compact regular expression '([klrs]?e | [bg]?o | b?a | di)'.

We formulate our queries at first on lemma level to be sure that incorrectly tagged items will still be found. However, as we would like to explore the tagging quality of the morphemes appearing in our structures, we will repeat the queries on npos and tpos level.

4.2 Developing macros for querying SEPEDI2021

In Table 1, Prinsloo (2020: 323) describes sequences of morphemes and conditions for a number of negation forms. We repeat parts of this table here as Table 4 that shows a productive perspective, in other words, how should a specific mood, tense and polarity be formulated? Working with corpus queries, we need to change to the receptive perspective: how should a sequence occurring in a corpus be interpreted?

Mood	*Negation strategy*
3.1 Indicative	
3.1.1 Pres.	***ga*** **+ *subject concord* + verb stem** ending **-e**
3.1.2 Fut.	**subject concord + *ka se* + verb stem** ending **-e**
3.1.3 Past	**1. *ga se* + alternative concord + verb stem** **2. *ga se* + subject concord + verb stem** ending **-e** **3. *ga* + subject concord + a + verb stem** **4. *ga* + alternative concord + verb stem**
3.2 Situative	
3.2.1 Pres.	**subject concord + *sa* + verb stem** ending **-e**
3.2.2 Fut.	**subject concord + *ka se* + verb stem** ending **-e**
3.2.3 Past	**subject concord + *sa* + verb stem**
3.3 Relative	
3.3.1 Pres.	**subject concord + *sa* + verb stem** ending **-e + -go/-ng**
3.3.2 Fut.	**subject concord + *ka se* + verb stem** ending **-e + -go/-ng**
3.3.3 Past	**subject concord + *sa* + verb stem + -go/-ng**
3.4 Subjunctive	**subject concord + *se* + verb stem** ending **-e**
3.5 Habitual	**subject concord + *se* + verb stem** ending **-e**
3.6 Consecutive	**alternative concord + *se* + verb stem** ending **-e**
3.7 Infinitive	***go* + *se/sa* + verb stem** ending **-e**
3.8 Imperative	**1. *se* + verb stem** ending **-e** **2. *se ke* + alternative concord + verb stem**

Table 4: Mood and negation strategies (Prinsloo 2020: 323)

We utilize the descriptions of Table 4 for performing corpus queries, but there are challenges:

1. There are several identical sequences appearing in different moods (there is syncretism), see for example the negation of the future tense of Indicative (3.1.2) and Situative (3.2.2) or Subjunctive (3.4) and Habitual (3.5).
2. The infinitive (3.7) contains the highly ambiguous class prefix *go*. It would exceed the scope of this work to distinguish all infinitive class prefixes from the subject concords of class 15 and the locative classes. Therefore, we will count the syncretic cases where *go* appears separately.
3. The endings of verb stems are not described in a number of categories, we thus must use other, more detailed descriptions of the negation forms additionally to be able to precisely formulate our queries.
4. The available taggers do not differentiate between the different sets of subject and alternative concords; we must therefore also search our corpus on the levels of lemma or token, respectively even when trying to find tokens on "npos" or "tpos" level. Still, we will not be able to allocate some of our results to one specific mood (without a linguistic expert reading and interpreting all of the respective sentences).
5. Lastly, we need to be more precise in our descriptions, as we want to take transitive verbs into account that might be preceded by an object concord.

To solve issue 3 and 5, we rely on the morpheme sequences described in the PhD Dissertation of Faaß (2010), which was supervised by D.J. Prinsloo. These are based on the theoretical descriptions of Lombard et al. (1985), Louwrens (1991), and Poulos and Louwrens (1994) and thus identify more ways of negating. Exploring for example the negated future tense of the relative, Prinsloo (2020) provides one possibility: (1) subject concord + *ka* + *se* + verb stem (we presume that it ends in -*a*); using Faaß (2010) for comparison, there are two more strategies: (2) subject concord + *ka* + *se* + *tla/tlo* + verb stem ending in -*a* + -*go*/-*ng* and (3) subject concord + *ka* + *se* + *tlago/tlogo* + verb stem ending in -*a*. To find all possible forms, we will look for all of the described sequences, however, in our results, we will number them alongside the definitions of Prinsloo (2020) as shown in Table 4, so that a link to his article is established.

A typical token sequence describing the negation strategies for the indicative presence (ibid.) with the respective queries added, is described in Table 5.

Mood	*Negation strategy*	*CQP Queries*
3.1 Indicative		
3.1.1 Pres	***ga* + *subject concord* + *verb stem*** ending –e	**word:** [lemma ="ga"] [lemma="([klrs]?e|[bg]?o|b?a|di)"] [lemma="([lrs]?e|[bgm]?o|b?a|di)"]? [lemma=".+e"] **npos:** [(pos ="MNEG") & (lemma="ga")] [npos="CS.+"] [npos="OC.+"]? [(pos="V" & lemma=".+e"] **tpos:** [(pos ="MORPH") & (lemma="ga")] [npos="CS.+"] [npos="OC.+"]? [(pos="V" & lemma=".+e"]

Table 5: Transferring a negation description into CQP queries

We must be realistic: SEPEDI2021 is rather small and far from being representative of the language. We hence decide to only explore the most frequent verb in this corpus as a case in point, but in a reproducible way so that this exploration can be repeated on other OCWB-encoded corpora and with other verbs.

First, we produce a ranking list of all tokens tagged as V and find *feta* ((to) "pass"/"exceed"[9]) on a high-ranking position as the most frequent unambiguous verb form with 2,025 (npos)/2,026 (tpos) occurrences. Other verbs are more frequent, but at the same time ambiguous in terms of their POS, so there is a risk of them being incorrectly tagged. Taking all of feta's inflectional and derivational forms appearing in the corpus, we count their occurrences in all moods and tenses in order to get an overview of the forms that the verb appears in. A positive intermediate result is that all of them are annotated as "V" by both taggers.

The past form of the Indicative, for instance, is described by four different negation strategies in our Table 4 (Table 1 of Prinsloo 2020: 323). Therefore it might be of interest to lexicographers, which negation strategies are followed for this verb or for that matter any other verb. Since such queries are stored in text files as so-called "macros", they can be freely exchanged between researchers and re-used at any given time.

Future investigations regarding other verbs are made possible because our queries are furnished with a variable ($0 in the queries shown below) that will be replaced by the query processor with a regular expression describing any verb stem provided at the time of query.

Table 6 shows how a regular expression (RegEx) is built for the existing *fet-* forms for exemplification reasons.

Freq. of occ.	***Word form***	***Comments***	***Translation***	***Building a RegEx (ignore upper/lower case)***
		Indicative		
2,024	*feta*	active (-a)	pass, exceed	
85	*fete*	active (-e)	(must) pass, exceed	fet[ae]
27	*fetwa*	passive (-a)	is passed, exceeded	
5	*fetwe*	passive (-e)	(must) be passed, exceeded	fetw?[ae]
6	*fetana*	active reciprocal (-a)	pass, exceed each other	fet(w?[ae] | ana)
92	*fetile*	active perfect (-ile)	passed, exceeded	
6	*fetilwe*	passive perfect (-ile)	was passed, exceeded	fet(il | an)?w?[ae]
		Relative		
153	*fetago*	active (-go)	who/which pass(es), exceeds	
2	*fetang*	active (-ng)	who/which pass(es), exceeds	feta(go | ng)
81	*fetego*	active (-go)	who/which does not pass, exceed	
1	*feteng*	active (-ng)	who/which does not pass, exceed	fet[ae](go | ng)
7	*fetwago*	passive (-go)	who/which is passed, exceeded	
5	*fetwego*	passive (-ng)	who/which is not passed, exceeded	fetw?[ae](go | ng)
521	*fetilego*	active perfect (-go)	who/which passed, exceeded	
9	*fetileng*	active perfect (-ng)	who/which passed, exceeded	
9	*fetilwego*	passive relative (-go)	who/which was passed, exceeded	fet(il)?w?[ae](go | ng)
3,033				

Table 6: Generating regular expressions for the occurring word forms of *fet-*

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1698 (Article)

We still need to deal with the problem of syncretism: 3.2.2 (future tense of the situative) in Prinsloo's table (2020: 323) is identical to 3.1.2 (future tense of the indicative). This shows that we cannot distinguish the two moods by only looking at the token sequences described. As the situative is often preceded by the conjunction *ge* ("when"), we first explore the typical distance between *ge* and a following verb and find a maximum of 3 tokens appearing, one of which may never be punctuation. We therefore add the condition that the token *ge* for the negated form of the future indicative may not appear in up to 3 tokens preceding the described sequence (while, for the situative, such an occurrence of *ge* is defined as obligatory). We are however conscious of the fact that the problem may only be partially solved.

An exemplifying macro is IND-FUT-NEG(1) in Figure 1. It describes the negated future tense of the indicative on lemma level (all of the indicative forms are summarized in Table 3.19 of Faaß (2010)). The variable $0 will be replaced by a verb form (without ending) when starting the macro (where the ending is pre-defined). Results of the query will first be written into a sub-corpus called _IND-FUT-NEG. These matches are then counted on a lemma level and the resulting table is written into a file called ind-fut-neg.csv (see Figure 2).

```
MACRO IND-FUT-NEG(1)
set MatchingStrategy longest;
show -cpos;
_IND-FUT-NEG =
[lemma!="ge"]{0,3}                  # no ge/Ge should precede the sequence
# future tense negative
[lemma="(b?a|[klrs]?e|[bg]?o|di)"]  # CS
[lemma="ka"]                        # MORPH_neg
[lemma="se"]                         # MORPH_neg
[lemma="([gbm]?o|b?a|[ls]?e|di)"]? # possible OC
[lemma="$0e"];                      # verb stem ending in -e
cat _IND-FUT-NEG;
count by word  > "/Users/faassg/corpora/sepedi2021/work/ind-fut-neg.csv";
;
```

Figure 1: Macro IND-FUT-NEG(1) finding all negated future tense indicative forms of a specific verb stem

The result of the macro IND-FUT-NEG(1), processed with the regular expression "fetw?" (the question mark stands for optionality of the previous character, thus we describe the active and the passive form of future tense) is shown in Table 7. It should be noted that the original .csv file contains matches with up to three tokens preceding the verb (described by [lemma!="ge"]{0,3}). We find 5 different sequences (types), one verb form in altogether 11 occurrences.

CS	*OC*	*NEG*	*NEG*	*VERB*	*freq*
le		*ka*	*se*	*fete*	6
e		*ka*	*se*	*fete*	3
a		*ka*	*se*	*fete*	1
di		*ka*	*se*	*fete*	1
				Total	11

Table 7: Summarized results of the Macro "IND-FUT-NEG", run with the verb stem expressed as "fet(il)?w?"

5. Results for *feta*

5.1 Results of the macros

In section 4 above, we developed a regular expression describing all indicative forms of the verb *fet-* appearing in the SEPEDI2021 corpus: fet(il|an)?w?[ae]. However, we need to delete the verbal endings [ae], as they are already described by the macro (see Table 1). The active reciprocal form *fetana* does not appear in the corpus with any other endings, it is thus only queried in the respective moods, tenses and polarities that display verbs ending in *-a*.

Table 8 shows our results. As mentioned above, occurrences of *go* in the indicative verbal phrases might be infinitives, therefore they are counted separately. The same applies to *se* in the subjunctive/habitual that was queried simultaneously. The morpheme *se* might be a negation morpheme instead of a subject or an object concord. To avoid counting identical forms twice for the subjunctive/habitual, all cases where *se* appears are counted as negated forms of the verb.

mood/tense	*macro run*	*type of sequences found*	*found verb forms fet-*	*freq*	*freq of go*	*Total*
3.1 Indicative						
3.1.1. Pres	IND-PRES-POS["fetw?"]	37	w?a	471	1,532	2,003
	IND-PRES-NEG["fetw?"]	7	w?e	16	0	16
3.1.2. Fut	IND-FUT-POS["fetw?"]	4	-a	10	0	10
	IND-FUT-NEG["fetw?"]	5	-e	5	0	5
3.1.3 Past	IND-PERF-POS["fet?"]	16	-ilw?e	126	n.a.	126
	IND-PERF-NEG["fetw?"]	2	-a	9	0	9
3.2 Situative						
3.2.1 Pres	SIT-PRES-POS["fetw?"]	15	-a	53	n.a.	53
	SIT-PRES-NEG["fetw?"]	0				0
3.2.2 Fut	SIT-FUT-POS["fetw?"]	0				0
	SIT-FUT-NEG["fetw?"]	0				0
3.2.3 Past	SIT-PERF-POS["fet(il)?w?"]	6	-ile	13	n.a.	13
	SIT-PERF-NEG["fet(il)?w?"]	0				0

3.3 Relative						
3.3.1 Pres	REL-PRES-POS["fetw?"]	24	-w?ago/ -ang	154	n.a.	154
	REL-PRES-NEG"fetw?"]	11	w?ego	78	n.a.	78
3.3.2 Fut	REL-FUT-POS["fetw?"]	0				0
	REL-FUT-NEG"fetw?"]	0				0
3.3.3 Past	REL-PERF-POS["fetilw?"]	0				0
	REL-PERF-NEG"fetw?"]	1	-a	8	n.a.	8
mood/tense	*macro run*	*type of sequences found*	*found verb forms fet-*	*freq*	*freq of se*	*SUM*
3.4 / 3.5 Subjunctive and Habitual	SUBJ-HABIT-POS["fetw?"]	11	-w?e	30	53	11
	SUBJ-HABIT-NEG["fetw?"]	10	-w?e		53	53
3.6 Consecutive	CONSEC-POS["fetw?"]	0				
	CONSEC-NEG["fetw?"]	0				
3.7 Infinitive	see column "go"					
3.8 Imperative	IMP["fet"] (pos and neg)	0				0
	sum					2,631

Table 8: Summarized results of the occurrences of forms of *fet-* in the corpus

5.2 Corpus data for developing dictionary entries for *fet-*

To find data usable for a theoretical dictionary entry for *feta* based on our (non-representative) corpus data, we explore the forms found and the sub-corpora generated by the macros in more detail. As these data are too extensive to be shown completely in a contribution of this kind, we attempt to summarize them here. Again, it must be stressed that any interpretation must be conscious of Sepedi syncretism and the non-representativeness of the corpus.

5.2.1 General data on the occurring word forms of *fet-*

fet- is not very frequently passivized (59 passive voice forms versus 2,968 active forms) and only one derivation, namely *fetana* ("pass, exceed each other") appears in the corpus. When forming the relative, the ending *-go* is clearly preferred (764 occurrences) while its alternative ending, *-ng*, only occurs 12 times.

5.2.2 Data on the occurrences of *fet-* in the different moods

As it is typical for most verbs, the use of the indicative seems to be decidedly preferred (2,175 occurrences), while the number of occurrences of the relative (240) is significantly higher than that of the situative (43). The situative/habitual appears a few times, but none of the other moods can be found in the corpus.

Concerning negation strategies, we find the following data:

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1698 (Article)

1. The indicative in the perfect tense (3.1.3 in Table 4) allows for four ways of negation, in all 8 occurrences of this negated mood *feta* makes use of *ga* + alternative subject concord + verb stem.
2. The negated past tense of the situative was described by Prinsloo as subject concord + *sa* + verb stem. Faaß (2010), on the basis of Lombard et al. (1985: 149), describes three possible ways, of which one appears in the corpus: *a se a fetwa* (subject concord + *se* + alternative subject concord + verb ending in *-a*.

5.2.3 Data on the occurrences of *fet-* in the different tenses

The present tense dominates the occurrences (2,289) of *fet-*, while the past/perfect tense appears far less frequently (154). The future tense seems to be rather irrelevant in this corpus (26).

5.2.4 Data on the polarity of *fet-*

As is the case for most verbs, its positive form appears by far more frequently: 2,352 positive sequences appear versus 170 negated sequences. However, for the relative, 86 negated sequences stand against 154 positives. Not counting the fact that the corpus is rather small, such data could lead to the suspicion of specific semantics of the verb in the relative — an aspect that could be further explored.

5.2.5 Data on the transitivity of *fet-*

We do not have a full overview of the transitivity of the verb *fet-* because we only check for occurrences of the object concord which stands for a known object in the discourse. An object usually occurs after the verbal structure. However, occurring object concords give a clear indication that the verb may appear in a transitive reading.

fet- appears in the corpus with and without object concords. In the indicative positive of the present tense, for example, there are 1,713 occurrences found without an object concord, of these, 1,535 show a preceding *go* (pointing to a possible infinitive). We find 7 occurrences where a subject concord is followed by *a* before the verb in its base (active and passive) form appears. This *a*, as stated above, may either be interpreted as a tense morpheme (long form of the present tense) or as an object concord. We find 13 more occurrences of *go* as the first element of the sequence, followed by a morpheme that could be an object concord. Lastly, we are left with 276 sequences in which the verb undoubtedly follows a sequence of subject and object concord.

5.3 Open issues for *fet-*

Altogether, 3,033 word forms of *feta* should have occurred as part of the predefined sequences, but only 2,522 were found. While syncretism certainly leads to finding doublettes, there are also some sequences found that do not appear as defined in books written for language learners.

Following these books, the relative, for example, should only appear as *fetilego,* namely with the verbal ending *-ego* when preceded by the negative morpheme *sa* in the negated perfect tense of the indicative. Preceded by *sa*, it does not occur at all in our corpus, with *ka se*, we also do not find any occurrences in SEPEDI 2021. Hence, all 539 occurrences of *fetilw?ego* are part of other sequences that we cannot define on the basis of the given literature. It would exceed the scope of this contribution attempting to interpret these cases. However, for possible future work in collaboration with linguists, Table 9 shows the forms and their preceding items as they occur in the corpus.

pos -2	***pos-1***	***found verb forms fet-***	***SUM***
ye	*e*	*-ilego*	222
wo	*o*	*-ilego*	84
tše	*di*	*-ilego*	50
yeo	*e*	*-ilego*	41
se	*se*	*-ilego*	26
le	*le*	*-ilego*	22
ao	*a*	*-ilego*	14
leo	*le*	*-ilego*	12
tšeo	*di*	*-ilego*	7
a	*a*	*-ilego*	6
bao	*ba*	*-ilego*	6
bjo	*bo*	*-ilego*	5
ye	*e*	*-ileng*	4
yeo	*e*	*-ilwego*	4
lebaka	*le*	*-ilego*	3
tše	*di*	*-ileng*	3
mmalwa	*ye*	*-ilego*	2
seo	*se*	*-ilego*	2
woo	*o*	*-ilego*	2
yeo	*e*	*-ileng*	2
3	*e*	*-ilego*	1
ao	*a*	*-ilwego*	1
bangwe	*ba*	*-ilego*	1
bao	*ba*	*-ilwego*	1
bja	*bao*	*-ilego*	1
bošegong	*bjo*	*-ilego*	1
6	*tše*	*-ilego*	1
e	*e*	*-ilego*	1
go	*go*	*-ilego*	1
go	*go*	*-ilwego*	1

go	*tše*	*-ilego*	1
kgoro	*ye*	*-ilego*	1
lekgolo	*e*	*-ilego*	1
mabaka	*a*	*-ilego*	1
mabakeng	*a*	*-ilego*	1
mengwaga	*ye*	*-ilego*	1
mo	*go*	*-ilego*	1
pedi	*tše*	*-ilego*	1
tse	*di*	*-ilego*	1
tše	*di*	*-ilwego*	1
tšeo	*di*	*-ilwego*	1
yo	*a*	*-ilego*	1
Total			539

Table 9: Occurrences of *fetilw?ego* in the corpus

6. Results for part-of-speech sequences

The macros are executed in two additional modified versions where sequences were queried on the basis of npos and tpos. Whenever the POS-set category included more than one item though it is explicitly specified in the definitions given, the item is named on lemma-level in these macros. For example, when a negative form has to contain the negative morpheme *ga*, the constraint is formulated [npos="MORPHNEG" & lemma="ga"] or [tpos="MORPH" & lemma="ga"], because other items like *se* are also classified as MORPH(NEG). If the POS category contains only one member, for example the present tense morpheme *a*, the constraint is defined on the POS level only ([npos="MORPHPRES"]).

mood/tense	***polarity***	***queried word form RegEx***	***verbal ending (def. in macro)***	***Total found for lemma***	***Total found for npos***	***Total found for tpos***
3.1 Indicative						
3.1.1. Pres	pos	["fetw?"]	-a	2,003	213	638
	neg	["fetw?"]	-e	16	0	14
3.1.2. Fut.	pos	["fetw?"]	-a	10	10	8
	neg	["fetw?"]	-e	5	0	1
3.1.3 Past	pos	["fet"]	(il|etš)w?e	126	102	108
	neg	["fetw?"]	-a, -e	9	0	1
3.2 Situative						
3.2.1 Pres	pos	["fetw?"]	-a	53	58	64
	neg	["fetw?"]	-e	0	0	0
3.2.2 Fut	pos	["fetw?"]	-a	0	5	3
	neg	["fetw?"]	-e	0	0	0
3.2.3 Past	pos	["fet"]	-(il|etš)w?e	13	20	16
	neg	["fet"]	-(il|etš)w?e	0	0	0

3.3 Relative						
3.3.1 Pres	pos	["fetw?"]	-a(go|ng)	154	145	148
	neg	["fetw?"]	-e(go|ng)	78	0	75
3.3.2 Fut	pos	["fetw?"]	-a(go|ng)?	0	0	0
	neg	["fetw?"]	-a(go|ng)?	0	0	0
3.3.3 Past	pos	["fetilw?"]	-a, -e	0	0	0
	neg	["fetw?"]	-a, -e	8	0	0
3.4 / 3.5 Subjunctive and Habitual	pos	["fetw?"]	-e	11	31	27
	neg	["fetw?"]	-e	53	0	11
3.6 Consecutive	pos	["fetw?"]	-a	0	0	0
	neg	["fetw?"]	-e	0	0	0
3.8 Imperative	pos	["fet"]	-a(ng)?	0	0	0
	neg	["fet"]	-e(ng)?	0	0	0
				2,631		

Table 10: Searching on npos and tpos-level for *fet-* in the corpus

In order to evaluate the taggers, in a first run, the respective verb forms of *fet-* were queried as well. Table 10, repeating the totals of Table 8 of the queries of the lemma level (for comparative reasons) shows the results. Note that in the case of the situative, all conjunctions were permitted to appear ([n/tpos="CONJ"]). There were more matches (*ge, gore, ebile* etc.) occurring than were found for the situative queried on lemma-level (only *ge*).

Comparing the number of occurrences found with and without a POS constraint, it is quite clear that the tagging quality especially of the ambiguous morphemes is still a problem. Here, the finely grained "MORPH" definitions of the NCHLT tagger seem especially problematic: to distinguish MORPHNEG, MORPHPRES and TENSE reduces the number of cases and leads to problems when training a heuristic tagger. For the TreeTagger, the tag "MORPH" was chosen for all of the abovementioned morphemes because they all appear in similar positions, that is within a similar context. Because of the more coarse-grained label, the tool can find more occurrences of this type of token in the training phase and thus its precision is enhanced.

As a computational linguist, one would need to dig deeper into this evaluation, however for the purpose of this article we can summarize that querying on lemma level without using POS constraints might be the better option — until such time that the tagging quality of the ambiguous items is enhanced.

Finally, we tried the same macros again, now without a constraint on the verb root. Table 11 (columns "npos" and "tpos") shows the numbers of occurrences of all sequences finalized with tokens annotated as verbs (with their verbal endings as defined above for each mood, tense, and polarity). Again, we must assume a number of doublettes (see for example the high number of subjunctives/habituals identified), caused by the syncretism explained above. Others will be tagged incorrectly — all in all, we can however get a general indication of which moods, tenses and polarities appear more frequently in the corpus than others. We know that the results do not seem sufficient for highly ambiguous items, so lastly, the macros were again defined on a lemma level — however now using [npos="V"] as the only constraint on their final element (adding the necessary endings as above). Results are shown in column "lemma+npos".

As the tagging results do not seem sufficient for highly ambiguous items, lastly, the macros were again defined on lemma level — however now using [npos="V"] as the constraint on their final element.

mood/tense	*polarity*	*verbal ending (def. in macro)*	*npos*	*tpos*	*lemma+npos*	*Totals*
3.1 Indicative						
3.1.1. Pres	pos	-a	49,072	64,053	138,909	
	neg	-e	0	1,720	2,819	
3.1.2. Fut.	pos	-a	8,244	7,097	8,583	
	neg	-e	0	47	1,162	
3.1.3 Past	pos	(il|etš)w?e	13,207	20,755	14,959	
	neg	-a, -e	0	201	1,384	
						167,816
3.2 Situative						
3.2.1 Pres	pos	-a	9,805	7,174	6,522	
	neg	-e	0	430	513	
3.2.2 Fut	pos	-a	1,420	802	115	
	neg	-e	0	1	74	
3.2.3 Past	pos	-(il|etš)w?e	2,654	2,526	1,194	
	neg	-(il|etš)w?e	0	155	155	
						8,573
3.3 Relative						
3.3.1 Pres	pos	-a(go|ng)	21,934	22,029	24,568	
	neg	-e(go|ng)	0	1,496	1,574	
3.3.2 Fut	pos	-a(go|ng)?	1,234	1,221	1,556	
	neg	-a(go|ng)?	0	2	72	
3.3.3 Past	pos	-a, -e	21,934	22,029	24,387	
	neg	-a, -e	0	0	201	
						52,358
3.4 / 3.5 Subjunctive and Habitual	pos	-e	38,691	48,866	38,691	
	neg	-e	0	760	2,224	
						40,915
3.6 Consecutive	pos	-a	0	0	0	
	neg	-e	0	0	0	
3.8 Imperative	pos	a(ng)?	0	0	1	
	neg	e(ng)?	0	0	1	

Table 11: Searching on npos and tpos-level for all tokens annotated as verbs in the corpus

6.1 Data on the occurrences of verbs in different moods

Like in the case of *fet-*, the indicative dominates the field with (reading column lemma+npos) 167,816 occurrences. The relative occurs with 52,358 occurrences, while the situative is again on rank three with 8,573 occurrences.

6.2 Data on the occurrences of verbs in different tenses

174,905 present tense sequences are found (some of which might however be infinitives), followed by past/perfect tense with 42,280 occurrences. The third rank is reserved for the future tense (11,562).

6.3 Data on the polarity of verbs in general

259,485 of all moods appeared in the positive, while 10,179 sequences were negated. The relation in the relative mood between the positive and the negative polarity does not seem significant.

7. Summary and possible future work

In this contribution, we attempted to gain some insights into how a Sepedi corpus can be compiled and annotated, and how it may assist a lexicographer with exploring a specific verb as it is used in the language. Corpus data will also assist when sorting negations of Sepedi verbs in a dictionary according to the frequencies they appear in.

We chose the verb *fet-* as a case in point because it is an unambiguous verb occurring frequently in our corpus. The majority of its occurrences could be assigned to pre-defined moods, tenses and polarities. We found that this verb has intransitive and transitive uses, that it occurs in the passive, but only one of the many possible derivations appeared in our corpus. In the case of the relative, speakers of the language seem to prefer the ending *-go* instead of *-ng* which would be available, too.

Given a bigger and more representative corpus, one could inter alia explore derivations of this and other verbs, however this corpus is at least a starting point.

In addition to the lack of resources, we find three main challenges when switching from a prescriptive to a receptive perspective:

1. Syncretism is certainly the biggest problem when analysing morphology and/or syntax of Sepedi sentences. Language experts together with computational linguists could in future work closely together exploring these constellations in more detail in an attempt to find more indicators in texts helping to disambiguate. In the longer term, we could even try to re-define

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1698 (Article)

the modal system as it is always problematic — not only for learners of the language — to distinguish token sequences semantically when they are 100% identical.

2. For highly ambiguous bound morphemes, tagging corpora with POS should help with the disambiguation, but the tagging quality does still not seem sufficient for such items (maybe this is caused by inappropriate tagsets, too). Here, newer technologies, possibly deep learning as already implemented for example by Schmid (2019) might be of help.
3. When comparing grammar books and corpus data, we find constellations which were not explained or described in standard grammars. It is therefore necessary to explore the living language further and to adapt the grammar books following a descriptive approach.

All results of this work are reproducible since the SEPEDI2021 corpus consists of freely available data, and since this corpus is annotated with freely available tools. In view of the fact that it is compiled from sources generated by others, it may not be forwarded to other researchers because of legal reasons. The corpus queries described here are stored in macros that the author shares freely on request by other non-commercial researchers.

8. Endnotes

[1] URL: https://sadilar.org

[2] It would go beyond the scope of this article to show negation strategies for all verbs (the corpus is too small for this), however the corpus queries developed here are written so that they are utilizable for other verbs, too.

[3] See https://vlo.clarin.eu. The CLARIN VLO collects metadata about available resources and tools for language research.

[4] See https://sadilar.org. SADiLaR offers its own repository, but also reports its resources to CLARIN.

[5] See https://repo.sadilar.org/handle/20.500.12185/270?show=full for more details.

[6] See https://repo.sadilar.org/handle/20.500.12185/330?show=full for more details.

[7] The MBT tagger parameter file used for a demo show case tagger on the AFLAT pages by De Pauw and De Schryver (https://aflat.org/sothotag) is not available for download, and we did not find any other available taggers for Sepedi.

[8] Available at https://repo.sadilar.org/handle/20.500.12185/326 though not mentioned in the SADiLaR list of Sepedi tools provided by the repository.

[9] All translations in this paper are taken from the *Oxford School Dictionary: Northern Sotho and English*. Oxford University Press. 2007.

9. Bibliography

Dahl, Ö. 1979. Typology of Sentence Negation. *Linguistics* 17: 79-106.

De Vries, N., M. Davel, J. Badenhorst and W. Basson. 2014. A Smartphone-based ASR Data Collection Tool for Under-resourced Languages. *Speech Communication* 56(1): 119-131.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1698 (Article)

Eiselen, E. and M. Puttkammer. 2014. Developing Text Resources for Ten South African Languages. *Proceedings of the Workshop on Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era. LREC, 9th International Conference on Language Resources and Evaluation, Reykjavik, Iceland, May 26-31, 2014:* 3698-3703.

Evert, S. and A. Hardie. 2011. Twenty-first Century Corpus Workbench: Updating a Query Architecture for the New Millennium. *Proceedings of the Corpus Linguistics 2011 Conference, ICC Birmingham, 20–22 July 2011.* Birmingham: University of Birmingham.

Faaß, G. 2010. *A Morphosyntactic Description of Northern Sotho as a Basis for an Automated Translation from Northen Sotho to English.* Ph.D. Dissertation. Pretoria, South Africa: University of Pretoria.

Faaß, G. 2018. Lexicography and Corpus Linguistics. Fuertes-Olivera, P. (Ed.). 2018. *The Routledge Handbook of Lexicography:* 123-137. Oxon, UK: Routledge.

Faaß, G., U. Heid, E. Taljard and D. Prinsloo. 2009. Part-of-Speech Tagging of Northern Sotho: Disambiguating Polysemous Function Words. *Proceedings of the EACL2009 Workshop on Language Technologies for African Languages (AfLaT 2009), Athens, Greece, 31 March 2009:* 38-45.

Goldhahn, D., M. Sumalvico and U. Quasthoff. 2016. Corpus Collection for Under-resourced Languages with More than One Million Speakers. Soria, C. et al. 2016: *LREC 2016 Workshop: Collaboration and Computing for Under-resourced Languages: Towards an Alliance for Digital Language Diversity (CCURL 2016), Portorož, Slovenia, 23 May 2016:* 67-73.

Lombard, D., E. van Wyk and P. Mokgokong. 1985. *Introduction to the Grammar of Northern Sotho.* Pretoria: J.L. van Schaik.

Louwrens, L. 1991. *Aspects of Northern Sotho Grammar.* Pretoria: Via Afrika.

Poulos, G. and L. Louwrens. 1994. *A Linguistic Analysis of Northern Sotho.* Pretoria: Via Afrika.

Prinsloo, D.J. 2020. Lexicographic Treatment of Negation in Sepedi Paper Dictionaries. *Lexikos* 30: 321-345. doi: https://doi.org/10.5788/30-1-1610

Schmid, H. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. *Proceedings of International Conference on New Methods in Language Processing. Manchester, UK.*

Schmid, H. 1995. Improvements in Part-of-Speech Tagging with an Application to German. *Proceedings of the ACL SIGDAT-Workshop, Dublin, Ireland:* 47-50.

Schmid, H. 2019. Deep Learning-based Morphological Taggers and Lemmatizers for Annotating Historical Texts. *Proceedings of DATeCH,* May 2019, Brussels, Belgium.

Schmidt, H. and F. Laws. 2008. Estimation of Conditional Probabilities with Decision Trees and an Application to Fine-grained POS Tagging. Scott, D. and H. Uszkoreit (Eds.). 2008. *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), 18–22 August 2008, Manchester, UK. Vol. 1:* 777-784. Manchester: COLING.

Tognini-Bonelli, E. 2001. Corpus Linguistics at Work. *Studies in Corpus Linguistics*. Amsterdam/ Philadelphia: John Benjamins.

Appendix: NCHLT and TreeTagger Tagsets

Morpheme	*NCHLT tagger**	*TreeTagger*
Verbs		
auxiliary	VAUX	VAUX
copulative	VCOP	VCOP
others	V	V
Nouns		
regular	N01a, N02b, N01-N10, N14, N16-N18, NLOC	N.01a, N.02b, N.01-N.10, N.14, N.LOC
name of place	—	NPP
abbreviation	—	ABBR
Pronouns		
emphatic	PROEMP01-PROEMP10, PROEMPLOC, PROEMPPERS	PRO.EMP.01-PRO.EMP.10, PRO.EMP.14, PRO.EMP.LOC, PRO.EMP.PERS
possessive	PROPOSS02-PROPOSS10, PROPOSS14, PROPOSSPERS	PRO.POSS.01-PRO.POSS.10, PRO.POSS.LOC, PRO.POSS.PERS
quantitative	PROQUANT01-PROQUANT10, PROQUANT14, PROQUANTLOC	PRO.QUANT.01-PRO.QUANT.10, PRO.QUANT.14-PRO.QUANT.15, PRO.QUANT.LOC
question word	QUE	QUE
Adverbs	ADV	ADV
Adjectives	ADJ01-ADJ10, ADJ14, ADJLOC	ADJ.01-ADJ.10, ADJ.14-ADJ15, ADJLOC
Morphemes		
negative	MNEG	MORPH
future	MORPHFUT	MORPH
? (always: *sa*)	MORPHPER	MORPH
potential (.**ka*)	MORPHPOT	MORPH
present tense (*w?a*)	MORPHPRES	MORPH
infinitive (*go*)	INF	MORPH
aspectual prefix (*no*)	ASP	MORPH
tense marker	TENSE	—
Concords		
subject	CS01-CS10, CS14-CS15, CSINDEF, CSLOC, CSNEUT, CSPERS	CS.01-CS10, CS.14-CS.15, CS.INDEF, CS.LOC, CS.NEUT, CS.PERS
object	CO01-CO10, CO14, COPERS	CO.01-CO.10, CO.14-CO.15, CO.LOC, CO.PERS
possessive	CPOSS01-CPOSS10, CPOSS14-CPOSS17, CPOSSLOC	CPOSS.01-CPOSS.10, CPOSS.14-CPOSS.15, CPOSS.LOC
demonstrative	CD01-CD10 CD14-CD18 CDLOC	CDEM.01-CDEM.10, CDEM.14, CDEM.COP, CDEM.LOC
Conjunctions	CONJ	CONJ
Particles		
question	PARTQUE	PART
others	PART	PART

Interjections	INT	INT
Enumeratives	ENUM	ENUM
Ideophones	IDEO	IDEO
Numerals	RS	NUM
Ordinals	RS	ORD
Punctuation		
.?	ZE	
!	ZE!	
,;-:	ZM	
left brackets/quotes	ZPL	
right brackets/quotes	ZPR	
.?!,;:		$.
brackets, quotes		$"
/\-%&		$-
Others		
Abbreviation of *Morena, Mna.* (=Mister, Mr.)	RO	ABBR
guess: foreign language material, however a number of Sepedi names (N01A and NPP) are tagged as RV	RV	—

* A full description of the NCHLT tagset could not be found, hence only the categories appearing in the corpus are described by the author in this table.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1699 (Article)

Critical Lexicography at Work: Reflections and Proposals for Eliminating Gender Bias in General Dictionaries of Spanish

Pedro A. Fuertes-Olivera, *International Centre for Lexicography, University of Valladolid (Spain) and Department of Afrikaans and Dutch, University of Stellenbosch, South Africa (pedro@emp.uva.es)*
and
Sven Tarp, *Centre for Lexicographical Studies, Guangdong University of Foreign Studies, China, Centre for Lexicography, University of Aarhus, Denmark and Department of Afrikaans and Dutch, University of Stellenbosch, South Africa (st@cc.au.dk)*

Abstract: This study highlights the fact that dictionaries are ideological texts that are very influential, because millions of people regard them as sources of authority. It shows that existing general dictionaries of Spanish are so gender-biased that they contribute to the upholding of unfair situations, for example, by making women invisible and maintaining gendered traditions based on male-centred power and ideology. In order to avoid such an unfair situation, we introduce several new ideas regarding the question of language and gender. We also show how this can be put into practice in a dictionary portal that we are constructing at the time of writing this article. Therefore, this article offers several specific solutions with the aim of making women *lexicographically* visible, promoting the use of inclusive language in public and private discourse and eliminating gendered practices.

Keywords: INCLUSIVE LANGUAGE, DICTIONARY, SPANISH, IDEOLOGY, PROFESSION NOUNS

Opsomming: Kritiese leksikografie aan die werk: Gedagtes oor en voorstelle vir die uitskakeling van gendervooroordeel in algemene Spaanse woordeboeke. In hierdie artikel word daar beklemtoon dat woordeboeke ideologiese tekste is wat 'n groot invloed uitoefen, aangesien miljoene mense woordeboeke as gesaghebbende bronne beskou. Daar word aangetoon dat bestaande algemene Spaanse woordeboeke so bevooroordeeld is t.o.v. gender dat hulle 'n bydrae lewer tot die aanmoediging van onbillike situasies, byvoorbeeld, deur vroue onsigbaar te maak en deur gendertradisies, wat gebaseer is op manlikgesentreerde mag en ideologie, te handhaaf. Om so 'n onbillike situasie te vermy, stel ons verskeie nuwe idees rondom die taal- en genderkwessie voor. Ons toon ook aan hoe dit in 'n woordeboekportaal, wat tydens die skryf van hierdie artikel deur ons saamgestel word, toegepas kan word. Hierdie artikel bied dus verskeie spesifieke oplossings aan met die doel om vroue *leksikografies* sigbaar te maak, om die gebruik van inklusiewe taal in die openbare en private diskoers te bevorder en om genderpraktyke uit te skakel.

Sleutelwoorde: INKLUSIEWE TAAL, WOORDEBOEK, SPAANS, IDEOLOGIE, BEROEPSNAAMWOORDE

1. Introduction

Scholars typically analyse the linguistic relationship between gender and sexuality under the tenets of well-established linguistics approaches, such as conversation analysis, corpus-based and corpus-driven analyses, (critical) discourse studies, ethnography of communication, multimodal discourse studies, language ideological analysis, stylistics, and so on (Holmes and Meyerhoff 2003). In her review of the *Handbook of Language and Gender,* for instance, Pichler (2005: 637) indicates that most of the publications in the *Handbook* have abandoned the traditional sociolinguistic conceptualisation of gender as an independent social variable composed of two components, one male and one female. Instead, most authors "appear to take a broadly constructionist approach to gender, viewing gender as being accomplished in interaction rather than as a fixed category". This view has led to different types of analysis, for example, those focusing on socio-pragmatics view the categories of 'sex' and 'gender' as complex and intertwined and have argued that "differences between women's and men's use of language were best accounted for by attending to societal power relations" (Holmes and King 2017: 121). A large body of research supports the above conclusion (e.g. Tannen 1990, Fuertes-Olivera 1992, 2007, Velasco-Sacristán and Fuertes-Olivera 2006). Holmes and King (2017) indicate that research into the relationship between gender and language suggests that

— power (see Keating 2009) is dynamically constructed and exercised; in other words, different participants may exercise power in different ways and, therefore, "power is constantly being constructed, negotiated, maintained and re-asserted, as people interact" (Holmes and King 2017: 122);

— power is systemic, that is, the norms and expectations of the most powerful groups are taken for granted in most situations and imitated (see Wodak 1999, Holmes 2005);

— power is a central component of leadership, as research on workplace interactions has shown (see Holmes 2006, Mullany 2007, Baxter 2010).

This paper adds to the above body of research by analysing the role lexicographers have played in maintaining some of the abovementioned imbalance, and it will show that lexicographers can and must adopt a different approach with the aim of making language dictionaries inclusive. We first describe the theoretical framework of the dictionary as an ideological text, and present the concepts of power and ideology and their influence in existing general dictionaries of Spanish. After formulating the research questions we will address in this article, we will then go on to describe some existing general dictionaries of Spanish, discussing the significance of the findings for gender and lexicographic theory and dictionary practice. We will then present the *Diccionarios Valladolid-UVa,* a brand-new dictionary project that aims to use inclusive language, eliminate gender bias and respond to demands from current Spanish society, which, first of all, is asking for the feminisation of official texts and, second, is fighting against male-dominant approaches to day-to-day life.

Regarding the first issue, the Vice-president of the 2018 Spanish socialist government officially asked the Royal Spanish Academy, which is the editor of the *Diccionario de la Lengua Española* (*DLE*), to take a leading role in two main tasks. Firstly, it should prepare a set of guidelines for recommending the use of inclusive language in Spanish official texts, such as the Spanish Constitution. Secondly, it should adapt the *DLE* to inclusive language guidelines, for example, by including the feminine forms of profession nouns and avoiding the use of male generic terms.

It seems that the Vice-president's request aimed at advancing the fight against the Spanish machista society. On the one hand, the new texts and dictionary articles will make the feminist agenda more visible, by indicating that the political debate must also be the object of gender considerations. On the other hand, this request will highlight the role language plays in the construction of society. Both objectives will promote the de-genderisation of Spanish society, that is, Spanish entrenched gender stereotypes should be eliminated as soon as possible. Unfortunately, the response of the Royal Spanish Academy (RAE) has been so tame that it has only accepted the inclusion of some feminine words, for example, the word *presidenta* ('female president') and some doubles (e.g. *presidentes y presidentas* ('male and female presidents') in the texts, but with no real change in its *DLE* (*Informe sobre el buen uso del lenguaje inclusivo en nuestra carta magna* 2020).

We do not know the rationale for the response given to the Vice-president. This does not, however, concur with a lot of research, especially — and to name just a few examples — those championed by feminist linguists (e.g. Mills and Mullany 2011) on language and gender, by Halliday (1978) on the social interpretation of language meaning, and by Van Dijk (1984, 1987) on the connection between ideology and language. These, and many more authors, have defended the social role of language, which should be analysed not only in isolation, that is, taking into consideration its formal properties, but also in terms of its context, for example, by focusing on language as text and discourse. This is what we will do in the following sections.

2. The dictionary as ideological text

Dictionaries can be described in various ways, and one of them uses a textual approach. Dubois and Dubois (1971) and Frawley (1989), among others, have analysed the conceptualisation of dictionaries as text by focusing on the information structure of the lexicographic text. Fuertes-Olivera and Nielsen (2018: 17), for instance, observe that the textual structure of dictionaries is initially connected with their functions(s). In this regard, dictionaries compiled to help users searching for knowledge contain texts such as systematic introductions that provide their potential users with a description of the knowledge structure of a particular domain, for example, accounting (Fuertes-Olivera 2009, Niño Amo and Fuertes-Olivera 2017). They also indicate that, essentially, dictionaries are information tools that have been compiled to offer assistance in certain types of situation in which they are consulted. This includes communicative situations in which interlocutors engage in communicative acts and need help to successfully complete the tasks; for example, when authors write texts in their native language and when translators translate texts into or from a foreign language. Another type of situation, often referred to as cognitive, is when a person needs to acquire knowledge about something in general or specific knowledge about a particular matter. This could be when students consult dictionaries in order to widen their knowledge basis prior to lectures and when translators make a consultation in order to acquire knowledge about a subject field as a prerequisite for properly understanding source texts. Dictionaries consulted in these situations are collections of different examples of text (Bergenholtz, Tarp and Wiegand 1999: 1763) which contain varieties of data that support more than one lexicographic function, that is, to provide certain types of assistance to certain types of users in certain types of use situations (see Fuertes-Olivera 2018, Fuertes-Olivera and Tarp 2014).

Secondly, dictionaries are usually divided into sections that often represent different text types. These sections can be use-related if they contain data that help people to use the dictionary properly and to its fullest extent, such as user guides, while other sections are function-related, containing data that provide a service insofar as they satisfy lexicographically relevant needs, such as wordlists and dictionary articles. This division into sections is particularly evident in print dictionaries, which can be described as special types of books that are divided into a number of chapters; however, online dictionaries may also contain similar sections, which are in effect different web pages under specific dictionary websites. Each sectional text type found in dictionaries can be a potential source or target text. For instance, the print version of the *DLE* (2014) contains information sections (e.g. the name of academicians, including those in charge of the *DLE*), use-related sections, such as an explanation of the main changes introduced in the 23rd edition of the dictionary, and function-related sections, for example, each dictionary article.

Use-related sections contain data sets that may be termed generic, in the sense that many of these give general guidance about dictionaries and can be reused in other dictionaries. However, data sets in function-related sections are less likely to be generic, in that they are directly dependent on aspects such as the domains selected for consideration and the dictionary functions. This means that function-related sections in, say, specialised dictionaries, are likely to be more difficult to decode, transfer and encode than those in general-language dictionaries (Fuertes-Olivera and Nielsen 2018).

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1699 (Article)

Since dictionary articles can be regarded as texts, it is appropriate to look briefly at relevant text levels. Two levels stand out: headwords, that is, texts that are made up of headlines that introduce their (text) topics, and co-texts, such as 'definition', 'sentence example', 'equivalent', 'lexicographic note', etc., each of which describes the meaning, usage and possible restriction of each headword. In rudimentary dictionaries, dictionary articles have very few elements; for example, some specialised bilingual dictionaries have only headwords and equivalents. However, modern lexicography prioritises user needs in communicative and cognitive situations and, therefore, dictionary articles are now likely to contain more than two textual elements in order to provide help that can satisfy user needs (see Fuertes-Olivera 2018, Fuertes-Olivera and Tarp 2014: 48-57, Fuertes-Olivera and Nielsen 2018, Nielsen 2018). This is the case in all the general language dictionaries that will be analysed below.

Dictionary sections, then, can be subjected to different types of analysis. For instance, they can be studied in terms of the external social factors relating to them. This is, basically, what critical discourse analysis has done with many different text types. We will follow suit and will analyse function-related dictionary texts under the tenets of 'critical lexicography', a concept developed by Kachru (1995) to explain that "[l]exicography and its products, dictionaries, are never value-free, apolitical or asocial. Instead, they are subject to ideology, power and politics" (Chen 2019: 362).

Critical lexicography can be connected with the feminist movement, e.g. De Beauvoir (1949). For the purpose of this article, we can enumerate some changes that show the influence of this movement in English dictionaries: the creation of lemmas such as *chair*, *chairperson*, *police officer*, etc., which replace the generic uses of *chairman*, *policeman*, and so on; the new uses and definitions of *human race* and *humankind* replacing 'generic man'; the use of notes and comments against sexist and racist meanings; see, for example, **man** in *Lexico*; the use of 'singular they' instead of 'generic he'; and the publication of the *Feminist Dictionary* by Kramarae and Treichler (1985) (see also Baron 1986, Fuertes-Olivera 1992, Hidalgo Tenorio 2000). More recently, critical lexicography has morphed into several varieties, Critical Lexicographical Discourse Studies (CLDS) representing one of them. This

> rests on the assumption that lexicography is a recontextualizing practice and that the dictionary, as a recontextualized discourse, is closely associated with other social/discursive practices and a site where ideological and social struggle take place. As a recontextualized discourse, the dictionary does not simply replicate its source or just 'transport' meaning; rather, it creates meaning; it rewrites and represents things in new ways (Chen 2015).

Under the tenets of CLDS, we will show that existing lexicographic practices have embedded their lexicographic data in their social context, have taken a political stance explicitly and have not offered their users ways of emancipating themselves from traditional forms of domination (Wodak and Meyer 2015).

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1699 (Article)

The following analysis and critique of general dictionaries of Spanish is based on the concepts of power and ideology. Power is exercised by dominant groups with the aim of exerting "domination, coercion and control of subordinate groups" (Simpson and Mayr 2010: 2). Ideology sustains the interests of groups by promoting a "coherent and relatively stable set of beliefs and values" (Wodak and Meyer 2015: 30). Both concepts are relevant because ideology is primarily transmitted and enacted through language and language also contributes to exercising and maintaining power. Furthermore, dictionaries describe languages in terms of lexicographers' particular ways of seeing the world and their social contexts. Hence, dictionaries are powerful tools for transmitting power and ideology. Very often, Spaniards are 'informed' that something is as it is because this is what the *DLE* says. In other words, the dictionary of the Royal Spanish Academy is acclaimed as the source of authority, although it was conceptualised more than 300 years ago and has not incorporated most of the social changes that have occurred during this time. Some researchers — e.g. Nissen (1986), Fuertes-Olivera (1992), Forgas Berdet (1996), Calero Vaquera (2010), Rodríguez Barcia (2012), and Cabeza Pereiro and Rodríguez Barcia (2013) — have argued that the *DLE* should be totally updated; for example, by eliminating any trace of sexism in its structures. One possible adaptation is connected with the lemmatisation policy of dictionaries, as we will show below.

For reasons of space, we will analyse some function-related texts that clearly show how existing general dictionaries of Spanish influence the maintenance of a male-dominant society in Spain and the Spanish-speaking world. We will analyse in particular three research questions:

1. To what extent are women visible in general dictionaries of Spanish?
2. If not, which function-related texts are contributing most to the (in)visibility of women?
3. Are there any proposals we should advocate for increasing the visibility of women, and thus for eliminating the gender bias in general dictionaries of Spanish?

3. Methodology

Bosque and Barrios Rodríguez (2018) show that the Spanish lexicographic market is dominated by four printed general language dictionaries, all of which have published new editions in the last 15 years (two of them also have retro-digitised versions): *Diccionario de uso del español actual* (Clave 2004; this was retro-digitised in 2010), *Diccionario de uso del español* (María Moliner 2007), *Diccionario del español actual* (Seco, Andrés and Ramos 2011), and *Diccionario de la lengua española* (*DLE* 2014; this was also retro-digitised in 2014).

Due to limitations of space, we will investigate the lexicographic treatment of 'profession nouns', that is, nouns referring to male and female professionals. Spanish official statistics show that there are 10.8 million men and 9.1 million women working in Spain in 2020, and that they are evenly distributed in socio-

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1699 (Article)

economic sectors such as teaching, health, justice and the civil service (Ministerio de Trabajo 2020). If there is a 50% chance of finding either a woman or a man working in these sectors, it can be hypothesised that existing dictionaries, especially those that have recently published new editions, should offer a balanced analysis of some of the common nouns used for referring to them in these four socioeconomic sectors:

— 'teaching': we will analyse the Spanish words *profesor*, *profesora* ('professor'), *maestro*', *maestra* ('teacher'), and *enseñante* ('teacher');
— 'health': we will analyse the Spanish words *médico*, *médica* ('doctor'), *enfermero*, and *enfermera* ('nurse');
— 'justice': we will analyse the Spanish words *juez*, *jueza* ('judge'), *abogado*, *abogada* ('lawyer'), *fiscal*, and *fiscala* ('public prosecutor'); and
— 'civil service': we will analyse the Spanish words *funcionario*, *funcionaria* ('civil servant'), *director*, *directora* ('director') and *auxiliar* ('assistant').

These words have been selected for three reasons. Firstly, they are very common words [e.g. they frequently appear in the *Corpus de referencia del español actual* (CREA) and the *Corpus del español del siglo XXI* (CORPES XXI)]. For instance, a search for the Spanish words *profesor* and *profesora* in CORPES XXI retrieves more than 45,000 and 5,000 hits, respectively. This difference is determined by the policy of lemmatisation used in the corpus, which is the same as the one used in general dictionaries of Spanish, as we will show below (e.g. Table 1 and Discussion).

Secondly, as common and very easy-to-find words, their lexicographic treatment will easily show whether or not power and ideology are still dominant in mainstream Spanish culture, as we must insist that Spaniards typically resort to the authority of dictionaries, especially that of the Royal Spanish Academy, when they are discussing public matters.

Thirdly, the selected words, which, to the best of our knowledge, have never been analysed from a sociolinguistic perspective, are not associated with relevant male or female features. In other words, there is no objective reason for treating them differently from a lexicographic point of view. If their lexicographic consideration reproduces a bias, it may mean that these general dictionaries of Spanish continue playing a role in the upholding of a gendered society.

Our analysis will focus on two function-related texts: headwords and definitions. Headwords — also called lemmas, dictionary words, or dictionary entries — are typically subjected to a process called lemmatisation. This allows lexicographers to group together the inflected forms of a word. For instance, the forms *eat*, *eats*, *ate*, *eating* and *eaten* are lemmatised under **eat**, which is the canonical form one may look up in a dictionary, as it represents the whole inflection paradigm.

The concept of definition has been the subject of scrutiny in different fields, e.g. Philosophy, Logic, Law, Linguistics, Terminology and Lexicography. For the purpose of this paper, definitions describe the meaning of the headword, that is, the "set of conditions which must be satisfied by a lexical unit in order to denote the extra-linguistic reality/ies which correspond(s) to each of its senses" (Fuertes-Olivera and Arribas Baño 2008: 69). Hence, they refer to the "specific set of data that explains the meaning of a lemma and which is clearly addressed to the lemma" (Nielsen 2011: 202).

4. Data and results

Headwords are usually selected depending on etymology, grammar, such as part-of-speech, frequency and traditions which are taken for granted. By way of example, the *DLE* selects two Spanish headwords for the Spanish word *alma*, one deriving from the Latin *anima* ('soul') and one from the Hebrew *almá* ('virgin'), and two headwords for the Spanish word *cantar*, one for a noun ('poem') and one for a verb ('sing'). In general, selections based on etymology and grammar have little, if any, trace of power and ideology.

Lemma selection based on taken-for-granted traditions and frequency, however, may be strongly influenced by power and ideology. The use of 'frequency' for selecting headwords has been gaining momentum since 1987, that is, after Sinclair and his team published the *Collins Cobuild English Language Dictionary* (Sinclair 1987), which paved the way for using corpora in dictionary-related activities. Although some lexicographers have insisted on the benefits of a corpus approach to lexicography (Hanks 2012), its use has not been able to avoid the influence of power and ideology in the selection process. For instance, almost all corpus-based dictionaries rely exclusively or to a large extent on written texts; in other words, the prioritisation of the written language over spoken language is ideological in nature, as favouring *langue* over *parole* is a rhetoric of standardisation which serves the "transmutation of standard language into mythical national languages" (Fairclough 1989: 22). The abovementioned general dictionaries of Spanish have not selected their headwords on frequency counts. Instead, they have continued using taken-for-granted traditions, three of which are relevant for this article: (a) lexicographers mainly (sometimes exclusively) select the lemma list from standard sources, typically from literary works and newspapers; (b) lexicographers standardise their lemma list by assuming that their generalisations include women and men and that these are socially and politically neutral; (c) space constraints demand the use of 'dictionarese', that is, typical dictionary conventions.

One of these conventions is the use of headwords such as **profesor, a**. This headword does not exist in natural Spanish (Table 1). It shows that lexicographers have taken for granted that a headword such as **profesor, a** is the lemma for the Spanish words *profesor* ('male teacher'), *profesora* ('female teacher'), *profesores* ('male teachers'), and *profesoras* ('female teachers'). In other words, headwords such as **profesor, a; profesor -a; profesor -ra;** and **profesor, ra** (Table 1) are lexicographic conventions, that is, *dictionarese*, which do not exist in running texts. This convention seems to rest on the idea that female nouns derive from male nouns by adding a final *a* (which sometimes happens but not always) and that users must know the different formulae employed in each dictionary, as the four dictionaries under analysis show (they are the following: **profesor, a; profesor -a; profesor -ra;** and **profesor, ra**). In some cases, they must know that they may have to add or eliminate one or more letters in order to create an authentic feminine word from the headword. For instance, a user looking up the headword **profesor, ra** (*DLE*) must eliminate *one r* to create the feminine word *profesora*. This explains why, in Tables 1 to 5, words such as *profesora* ('female teacher'), *maestra* ('female teacher'), *médica* ('female doctor'), *enfermera* ('female nurse'), *jueza* ('female judge'), *abogada* ('female lawyer'), *fiscala* ('female public prosecutor'), *funcionaria* ('female civil servant') and *directora* ('female director') are either absent as headwords or, when present, are described as the wife of a man who may be a doctor, a judge, a civil prosecutor, and so on.

Words (profession nouns)	**Headwords**			
	Clave	*María Moliner*	*Seco et al.*	*DLE*
profesor ('male professor')	**profesor, a**	**profesor, -a**	**profesor –ra**	**profesor, ra**
profesora ('female professor')	x	x	x	x
maestro ('male teacher')	**maestro, tra**	**maestro, a**	**maestro –tra**	**maestro, tra**
maestra ('female teacher')	x	x	x	x
nseñante ('teacher'; 'professor')	**enseñante**	**enseñante**	**enseñante**	**enseñante**
médico ('male doctor')	**médico, ca**	**médico, -a**	**médico –ca**	**médico, ca**
médica ('female doctor')	x	x	x	**médica**
enfermero ('male nurse')	**enfermero, ra**	**enfermero, -a**	**enfermero, -ra**	**enfermero, ra**
enfermera ('female nurse')	x	x	x	x
juez ('male judge')	**juez**	**juez, -a**	**juez –za**	**juez, za**
jueza ('female judge')	**jueza**	x	x	**jueza**
abogado ('male lawyer')	**abogado, da**	**abogado, a**	**abogado –da**	**abogado, da**
abogada ('female lawyer')	x	x	x	x
fiscal ('male public prosecutor')	**fiscal**	**fiscal**	**fiscal –la**	**fiscal, la**
fiscala ('female public prosecutor')	**fiscala**	**fiscala**	x	**fiscala**
funcionario ('male civil servant')	**funcionario, ria**	**funcionario, a**	**funcionario -ria**	**funcionario, ria**
funcionaria ('female civil servant')	x	x	x	x
director ('male director')	**director, -a**	**director, -a**	**director –triz**	**director, ra**
directora ('female director')	x	x	x	x
auxiliar ('assistant')	**auxiliar**	**auxiliar**	**auxiliary**	**auxiliar**

Table 1: Headwords for some common profession nouns in certain selected general dictionaries of Spanish

Regarding definitions, Rundell (2015: 314) indicates that, in the printed era, a focus on economy led to definitions "which achieve conciseness (and aspire to precision) through the use of standard formulae ('the act of X-ing), 'characterised by Y', and so on) and through a recursive strategy". These strategies have costs that are passed on to the user, who has to learn these conventions in order to understand what the dictionary is saying (e.g. the previously mentioned Spanish headword **profesor, a**). He adds that in the last 30 years publishers, and especially those in the UK, have addressed this issue by developing more open defining styles. These aim to offer enough information for understanding the definition without knowledge of 'dictionarese', that is, the typical dictionary conventions such as the use of a recursive strategy, always assuming that

> a lexicographical definition (...) does not identify a meaning independently existing in actual usage and *discovered* there by the lexicographer: it is deliberately *constructed* and *allocated* by the lexicographer on the basis of materials selected for study, and its allocation will depend on the viewpoint the lexicographer has chosen to adopt. (Harris and Hutton 2007: 78)

and

> A definition can only be as effective as the context allows it to be, and the context includes the situation of the person seeking to understand the meaning. The notion of a definition adequate to all occasions and all demands is a semantic *ignis fatuus*. (Harris and Hutton 2007: 49)

Tables 2 to 5 show the definitions used for each of the headwords included in the dictionaries under analysis. We will include only the definitions of the nouns referring to the women and men employed in the abovementioned socioeconomic sectors. We assume that if these are evenly distributed, both women and men will be clearly identified. Some of the definitions are accompanied by lexicographic notes that are relevant for the topic under examination in this article.

Headwords	Definition (+ notes)
profesor, a	Persona que se dedica a la enseñanza, esp. si esta es su profesión (A high school teacher or university academic)
maestro, tra	1. Profesor de educación infantil o primaria (A person who teaches, especially in a nursery or primary school) 2. Persona que enseña una ciencia, un arte o un oficio, esp. si está titulada para ejercerlo. (A person who holds a degree for teaching a particular skill, profession, etc.)
enseñante	Referido a una persona, que se dedica profesionalmente a la enseñanza. (Any person who is professionally involved in teaching)
medico, ca	Persona legalmente autorizada para ejercer la medicina. (A person who is qualified to treat people who are ill)

enfermero, ra	Persona que se dedica profesionalmente a la asistencia de enfermos y heridos, esp. la que actúa como ayudante del médico. (A person trained to care for the sick or infirm, especially a doctor's assistant)
juez	Persona legalmente autorizada para juzgar, sentenciar y hacer ejecutar la sentencia. (A public officer appointed to decide cases in a law court)
jueza	➔ juez
abogado, da	Persona legalmente autorizada para defender a sus clientes en los juicios o aconsejarlos sobre cuestiones legales. (A person who practises law, e.g. by representing clients in a court, by advising clients on legal matters)
fiscal	(En algunas zonas del español meridional se usa el femenino *fiscala*). Persona legalmente autorizada para acusar de los delitos ante los tribunales de justicia. (In some areas of Spain, the feminine public prosecutor 'fiscala' is used instead of the masculine public prosecutor 'fiscal). (A law officer who conducts criminal proceedings on behalf of the state or in the public interest)
fiscala	➔ fiscal
funcionario, ria	Persona que desempeña un empleo en uno de los cuerpos de la Administración pública. (A member of the civil service)
director, -a	Persona a cuyo cargo está la dirección de algo. (A person who is in charge of something)
auxiliar	Referido a una persona, que ayuda o colabora en las funciones de otra como subordinada suya. (A person who ranks below a senior person)

Table 2: Definitions of the headwords in *Clave (our translations or glosses in brackets)*

Headwords	Definition (+ notes)
profesor, -a	Persona que enseña una determinada materia. (A person who teaches a specific subject)
maestro, -a	1. En sentido amplio, persona que enseña cualquier cosa, generalmente con respecto a quien recibe la enseñanza. (In a generic sense, a person who teaches anything, generally with respect to the one who is taught) 2. En sentido restringido, persona que, con o sin título oficial para ello, da la primera enseñanza. (In a restricted sense, a person who teaches in a primary school, who may or may not have a degree)

enseñante	Que enseña (hace que alguien aprenda). (Someone who teaches) (they help someone learn something)
medico, -a	Persona que tiene título oficial para curar las enfermedades. (A person who is qualified to treat people who are ill)
enfermero, -a	Persona que atiende a los enfermos en los hospitales, clínicas, etc., y ayuda a los médicos. (A person who cares for the sick in hospitals and assists doctors)
juez, -a	En sentido restringido, funcionario encargado de administrar justicia o decidir quién tiene razón en un pleito, en los tribunales públicos. (In a restricted sense, a public officer appointed to decide cases in a law court)
abogado, -a	Persona que tiene la carrera de derecho. Persona con esa Carrera que aconseja en asuntos de derecho o interviene en los juicios y procesos representando a una de las partes. (A person who holds a degree in law. A person who holds a degree in law and gives advice on legal matters and acts in a case on behalf of one of the parties involved)
fiscal	Funcionario de la Carrera judicial que representa en los juicios el interés público y, con ese carácter, mantiene la acusación contra los delincuentes o el interés del Estado frente a los particulares, en contra del abogado defensor de éstos. (A law officer who conducts criminal proceedings on behalf of the state or in the public interest)
fiscala	Mujer que ejerce el cargo de fiscal. (A female law officer who conducts criminal proceedings on behalf of the state or in the public interest)
funcionario, -a	Empleado que está al servicio de la administración pública. (A member of the civil service)
director, -a	Persona encargada de dirigir cierta cosa. (A person in charge of managing something)
auxiliar	Funcionario en la categoría inferior en los cuerpos administrativos. (A civil servant enrolled in the lower ranks of the civil service)

Table 3: Definitions of the headwords in *María Moliner (our translations or glosses in brackets)*

Headwords	Definition (+ notes)
profesor –ra	Pers. que enseña [una ciencia o arte]. (A person who teaches science or arts)
maestro –tra	*Sin compl*: Pers. que tiene título oficial para enseñar en una escuela primaria. (A person who holds a degree for teaching in a primary school)
maestra	➔ maestro
enseñante	Que enseña (One who teaches)
medico –ca	1. Pers. que ha hecho la carrera de medicina. (A person who holds a degree in medicine) 2. Mujer del médico (A doctor's wife)
enfermero –ra	1. Pers. que se dedica profesionalmente a cuidar enfermos bajo las órdenes del médico. (A person who takes care of sick people under a doctor's supervision) 2. Pers. que cuida a un enfermo. (A person who takes care of sick people)
juez –za	(*La forma JUEZ se usa como m y f en accepts 1 y 2; la forma f JUEZA solo en accept 1)* (*The form JUEZ is used as masculine and feminine in meanings 1 and 2; the feminine form JUEZA is only used in meaning 1*) 1. Letrado con autoridad para juzgar y sentenciar. (A person who holds a degree in law and has authority for deciding cases in a law court)
abogado –a	(*a veces en acep 1, se usa la forma m con valor f*) (Sometimes in meaning 1 the masculine form is used as if it were feminine) 1. Licenciado en Derecho. (A person who holds a degree in Law) 2. Licenciado en Derecho que en un juicio o un proceso defiende [a una de las partes]. (A person who holds a degree in Law and he or she defends one of the parties involved in court) 3. Licenciado en Derecho que asesora [a alguien] (*comp de posesión*) en asuntos legales. (A person who holds a degree in law and advises someone in legal matters)
fiscal –la	(*la forma FISCALA solo en acep 4, donde gralm se usa la forma FISCAL como f*) (The form FISCALA only in meaning 4, where FISCAL is generally used as the feminine form) 4. En un juicio: Acusador público. (In a lawsuit, public prosecutor)

funcionario ria	1. Pers. que ocupa como titular un empleo en la función pública. (A person who works as civil servant)
director –triz	(*f DIRECTORA en acep 4*) (DIRECTORA in meaning 4) 4. Pers. encargada de dirigir. (A person who manages something)
auxiliar	Funcionario o empleado, técnico o administrativo, de categoría inferior. (An assistant civil servant, employee, technician or administrative worker)

Table 4: Definitions of the headwords in *Seco et al. (our translations or glosses in brackets)*

Headwords	Definition (+ notes)
profesor, ra	Persona que ejerce o enseña una ciencia o arte. (A person who practices or teaches science or arts)
maestro, tra	1. Persona que enseña una ciencia, arte u oficio, o tiene título para hacerlo. (A person who teaches science, arts, occupational skills, or holds a degree for teaching) 2. maestro de primera enseñanza (teacher in a primary school) 3. Hombre que tenía el grado mayor en filosofía, conferido por una universidad (A man who holds the highest degree in philosophy, given by some universities)
maestra	coloq.p.us. Mujer del maestro (colloquial; not much used; teacher's wife)
enseñante	Que enseña. U.t.c.s. (Someone who teaches; also used as a noun)
medico, ca	Persona legalmente autorizada para ejercer la medicina. (A person who is legally authorized to work as a doctor)
médica	1. Coloq. desus. Mujer del médico. (Colloquial. Not used. Doctor's wife)
enfermero, ra	Persona dedicada a la asistencia de los enfermos. (A person who gives assistance to sick people
juez, za	Para el f. u.t. la forma *juez* en acceps. 1-3) (For the feminine, please use the masculine form *juez* for meanings 1–3) 1. Persona que tiene autoridad y potestad para juzgar y sentenciar. (A person who has the authority to decide cases in a court of law) 2. Miembro de un jurado o tribunal. (A member of a jury or a board)

	3. Persona nombrada para resolver cualquier asunto o materia, especialmente una duda o controversia. (A person appointed for deciding on any issue, especially a doubt or controversy)
jueza	Coloq.p.us. Mujer del **juez**. (Colloquial. Not much used. Judge's wife)
abogado, da	1. Licenciado en derecho que ofrece profesionalmente asesoramiento jurídico y que ejerce la defensa de las partes en los procesos judiciales o en los procedimientos administrativos. (A person who holds a degree in law and gives professional advice as well as assisting in a court of law or in administrative proceedings) 2. Intercesor o mediador (Intercessor or mediator)
fiscal, la	La forma *fiscala* u. solo en aceps. 2 y 7; para el f., u.m. *fiscal* en acep. 2. (The form *fiscala* is used only in meanings 2 and 7; for the feminine it is better to use the masculine *fiscal* in meaning 2) 2. Persona que representa y ejerce el ministerio público en los tribunales. (A person who represents and practises as a public prosecutor) 7. Ver fiscala (See fiscala)
fiscala	Coloq. Desus. Mujer del **fiscal**. (Colloquial. Not used. Public prosecutor's wife)
funcionario, ra	Persona que desempeña profesionalmente un empleo público. (A person who works as a civil servant)
director, ra	Persona que dirige algo en función de su profesión o cargo. (A person who directs something depending on his or her position)
auxiliar	1. En los ministerios y otras dependencias del Estado, funcionario técnico o administrativo de categoría subalterna. (In ministries and other government departments, an assistant civil servant) 2. Profesor encargado de sustituir a los catedráticos en ausencias y enfermedades (A professor who substitutes for full professors for whatever reason)

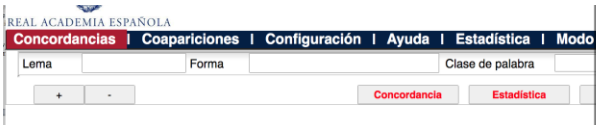
Table 5: Definitions of the headwords in *DLE (our translations or glosses in brackets). We are using the online version)*

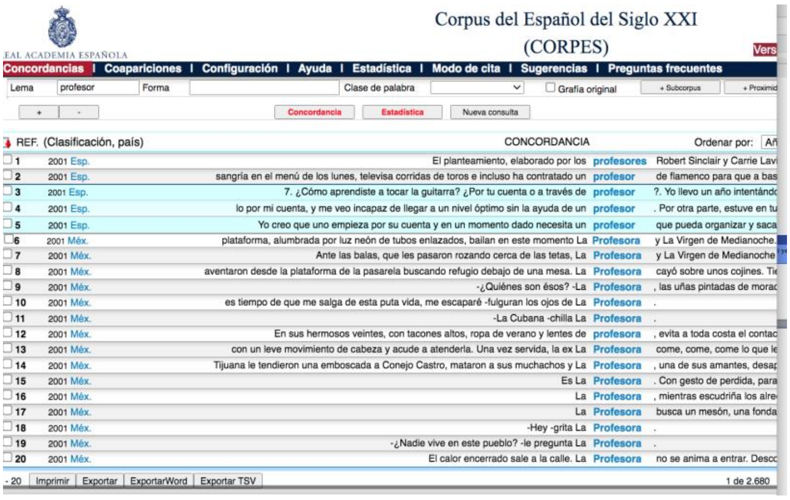
Tables 1 to 5 show three main results that will be discussed below. Firstly, Spanish dictionaries do not use natural words as headwords. Secondly, they lemmatise conventions that are typically formed by using the male form as the base form of the convention. Finally, women are either not specifically mentioned or, when mentioned, are usually treated as 'the wife of a professional man'.

5. Discussion and lexicographic solutions

From a qualitative point of view, Tables 1 to 5 show that women are mostly absent from general dictionaries of Spanish, and that when they are included, they are not presented fairly. Firstly, the use of headwords such as **profesor, ra** eliminates the figure of women from dictionaries. Although millions of Spaniards agree and take for granted that dictionaries are sources of authority, only a handful of them have received some training in 'dictionarese', and therefore will easily assume, say, that *ra* is a kind of symbol for the Spanish word *profesora*. For many Spaniards, 'dictionarese' such as **profesor, a** and **profesor, -ra** are meaningless because they are conventions that must be learned, e.g. at school. Unfortunately, these conventions are not taught at school. Hence, such headwords greatly contribute to the invisibility of women. Nothing can be more dangerous for the public image of a person than to make them invisible.

Secondly, dictionarese forms such as **professor –ra** are converted into 'masculine' forms straightaway. For instance, in CORPES XXI the search system of this corpus allows users to search for lemma and form (Figure 1). The Spanish word *profesor* is assumed to be the lemma or headword, whereas *profesora* is a form. Any search with *profesor* will retrieve all the tokens of the headword **profesor, ra** [e.g. *profesor*, *profesora*, *profesores*, and *profesoras*], whereas any search with the token *profesora* will only retrieve hits of *profesora* and *profesoras* as forms. In sum, this mechanism helps explain that the Spanish word *profesor* is 9 times more frequent in CORPES XXI than its feminine counterpart *profesora:* 2,680 hits against 338 (Figure 1):



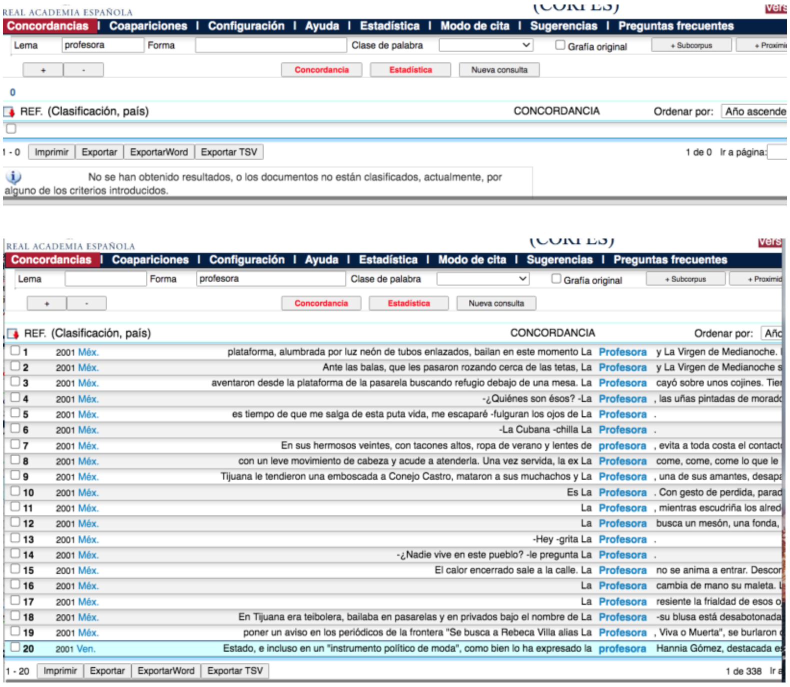


Lema profesor Forma

□ Grafía original

REF.	(Clasificación, país)	CONCORDANCIA		
□ 1	2001 Esp.	El planteamiento, elaborado por los	profesores	Robert Sinclair y Carrie Lavi
□ 2	2001 Esp.	sangría en el menú de los lunes, televisa corridas de toros e incluso ha contratado un	profesor	de flamenco para que a bas
□ 3	2001 Esp.	7. ¿Cómo aprendiste a tocar la guitarra? ¿Por tu cuenta o a través de	profesor	?. Yo llevo un año intentándc
□ 4	2001 Esp.	lo por mi cuenta, y me veo incapaz de llegar a un nivel óptimo sin la ayuda de un	profesor	. Por otra parte, estuve en tu
□ 5	2001 Esp.	Yo creo que uno empieza por su cuenta y en un momento dado necesita un	profesor	que pueda organizar y saca
□ 6	2001 Méx.	plataforma, alumbrada por luz neón de tubos enlazados, bailan en este momento La	Profesora	y La Virgen de Medianoche.
□ 7	2001 Méx.	Ante las balas, que les pasaron rozando cerca de las tetas, La	Profesora	y La Virgen de Medianoche
□ 8	2001 Méx.	aventaron desde la plataforma de la pasarela buscando refugio debajo de una mesa. La	Profesora	cayó sobre unos cojines. Ti
□ 9	2001 Méx.	-¿Quiénes son ésos? -La	Profesora	, las uñas pintadas de morac
□ 10	2001 Méx.	es tiempo de que me salga de esta puta vida, me escaparé -fulguran los ojos de La	Profesora	.
□ 11	2001 Méx.	-La Cubana -chilla La	Profesora	.
□ 12	2001 Méx.	En sus hermosos veintes, con tacones altos, ropa de verano y lentes de	profesora	, evita a toda costa el contac
□ 13	2001 Méx.	con un leve movimiento de cabeza y acude a atenderla. Una vez servida, la ex La	Profesora	come, come, come lo que le
□ 14	2001 Méx.	Tijuana le tendieron una emboscada a Conejo Castro, mataron a sus muchachos y La	Profesora	, una de sus amantes, desap
□ 15	2001 Méx.	Es La	Profesora	. Con gesto de perdida, para
□ 16	2001 Méx.	La	Profesora	, mientras escudriña los alre
□ 17	2001 Méx.	La	Profesora	busca un mesón, una fonda
□ 18	2001 Méx.	-Hey -grita La	Profesora	.
□ 19	2001 Méx.	-¿Nadie vive en este pueblo? -le pregunta La	Profesora	.
□ 20	2001 Méx.	El calor encerrado sale a la calle. La	Profesora	no se anima a entrar. Descc

- 20 1 de 2.680



Lema: profesora Forma: Clase de palabra: □ Grafía original

0

REF. (Clasificación, país) CONCORDANCIA Ordenar por: Año ascende

□

1 - 0 1 de 0 Ir a página:

No se han obtenido resultados, o los documentos no están clasificados, actualmente, por alguno de los criterios introducidos.

Lema: Forma: profesora Clase de palabra: □ Grafía original

REF. (Clasificación, país)			CONCORDANCIA		Ordenar por: Añ
□ 1	2001	Méx.	plataforma, alumbrada por luz neón de tubos enlazados, bailan en este momento La	Profesora	y La Virgen de Medianoche.
□ 2	2001	Méx.	Ante las balas, que les pasaron rozando cerca de las tetas, La	Profesora	y La Virgen de Medianoche s
□ 3	2001	Méx.	aventaron desde la plataforma de la pasarela buscando refugio debajo de una mesa. La	Profesora	cayó sobre unos cojines. Tie
□ 4	2001	Méx.	-¿Quiénes son ésos? -La	Profesora	, las uñas pintadas de morad
□ 5	2001	Méx.	es tiempo de que me salga de esta puta vida, me escaparé -fulguran los ojos de La	Profesora	.
□ 6	2001	Méx.	-La Cubana -chilla La	Profesora	.
□ 7	2001	Méx.	En sus hermosos veintes, con tacones altos, ropa de verano y lentes de	profesora	, evita a toda costa el contact
□ 8	2001	Méx.	con un leve movimiento de cabeza y acude a atenderla. Una vez servida, la ex La	Profesora	come, come, come lo que le
□ 9	2001	Méx.	Tijuana le tendieron una emboscada a Conejo Castro, mataron a sus muchachos y La	Profesora	, una de sus amantes, desapa
□ 10	2001	Méx.	Es La	Profesora	. Con gesto de perdida, parad
□ 11	2001	Méx.	La	Profesora	, mientras escudriña los alred
□ 12	2001	Méx.	La	Profesora	busca un mesón, una fonda,
□ 13	2001	Méx.	-Hey -grita La	Profesora	.
□ 14	2001	Méx.	-¿Nadie vive en este pueblo? -le pregunta La	Profesora	.
□ 15	2001	Méx.	El calor encerrado sale a la calle. La	Profesora	no se anima a entrar. Descor
□ 16	2001	Méx.	La	Profesora	cambia de mano su maleta.
□ 17	2001	Méx.	La	Profesora	resiente la frialdad de esos o
□ 18	2001	Méx.	En Tijuana era teibolera, bailaba en pasarelas y en privados bajo el nombre de La	Profesora	-su blusa está desabotonada
□ 19	2001	Méx.	poner un aviso en los periódicos de la frontera "Se busca a Rebeca Villa alias La	Profesora	, Viva o Muerta", se burlaron
□ 20	2001	Ven.	Estado, e incluso en un "instrumento político de moda", como bien lo ha expresado la	profesora	Hannia Gómez, destacada e

1 - 20 1 de 338 Ir a

Figure 1: Homepage of the *Corpus del Español del Siglo XXI* (CORPES XXI) and search system for 'profesor' and 'profesora'

Thirdly, Spanish lexicographers are making women invisible subconsciously, for example, by creating software that transforms headwords such as **profesor, ra** into masculine forms (*profesor*) (see the search mechanism in CORPES XXI in Figure 1), and by creating male and well-differentiated headwords for referring to powerful males. For instance, for more than a thousand years Spaniards have used the feminine word *modista* for referring to a person whose job is making clothes, typically women's dresses. As soon as men gained status in the profession, Spanish lexicographers created the masculine headword **modisto**, which corresponds to the actual word *modisto* ('couturier'). Furthermore, they defined it as 'a man who created women's dresses and fashion'. Why have Spanish lexicographers not created the headwords **modista, o** or **modista, to** for this new word is an open question, as many Spanish lexicographers usually claim that masculine words typically end in *o* in the same way that they assume that feminine words typically end in *a* (see headwords in Table 1, most of which have a final *a* as a kind of symbol for the feminine word). And, if dictionaries such as the *DLE* now lemmatise **modista** ('dressmaker') and **modisto** ('couturier') as headwords, why have they avoided the same lexicographic process with the Spanish feminine words *profesora*, *maestra*, *enfermera*, and so on?

To the best of our knowledge, nobody has answered this question, as shown below.

Fourthly, Spanish lexicographers use both rules that exist only in formal grammars as well as taken-for-granted traditions. These rules do not always work in language, especially in informal face-to-face encounters. Fuertes-Olivera (1992), for instance, has shown that Spaniards use social gender with some nouns (e.g. those relating to a profession). Social gender implies the use of male or female generics depending on users' social expectations and conventions. For instance, Spanish mainstream newspapers have recently used and continue using headlines such as *Las feministas se manifiestan contra la sentencia de la Manada* instead of *Los feministas se manifiestan contra la sentencia de la Manada* [the former is feminine and the latter masculine, and the headlines refer to the uproar caused by the rape of a young woman at the hands of a group of young men, identified as 'la Manada' ('the herd')]. Spanish newspapers use feminine generics (*Las feministas*) because Spanish society associates 'feminist' with women. This also explains why the generic expression *las feministas* is 42 times more used than *los feministas* in Google Books Ngram Viewer (Figure 2; see Pechenick, Danforth and Dodds (2015) for an analysis of some of the limitations of the Google Books Corpus). Such a figure would be impossible if the rules explaining the use of masculine words or expressions as generics were real. Instead, they are constructions based on the influence of power and ideology:

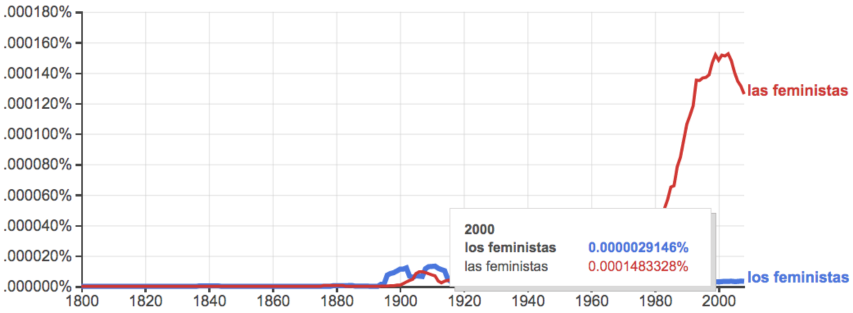


Figure 2: Rate of Frequency of *las feministas* and *los feministas* in Google Books Ngrams

Fifthly, Spanish lexicographers immediately created the headword **modisto** ('couturier') for referring to a man influencing fashion and making women's dresses. Why, then, have they not created the headword **enfermera** ('female nurse') if this word tends to be used 90% of the times in which Spaniards refer to the person trained to care for the sick or infirm, especially in hospitals (Figure 3):

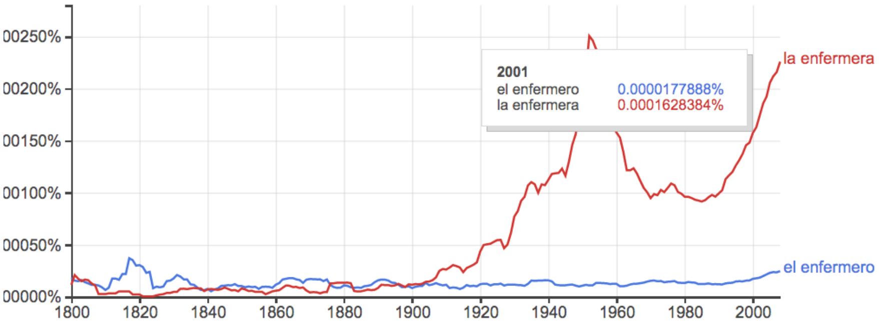
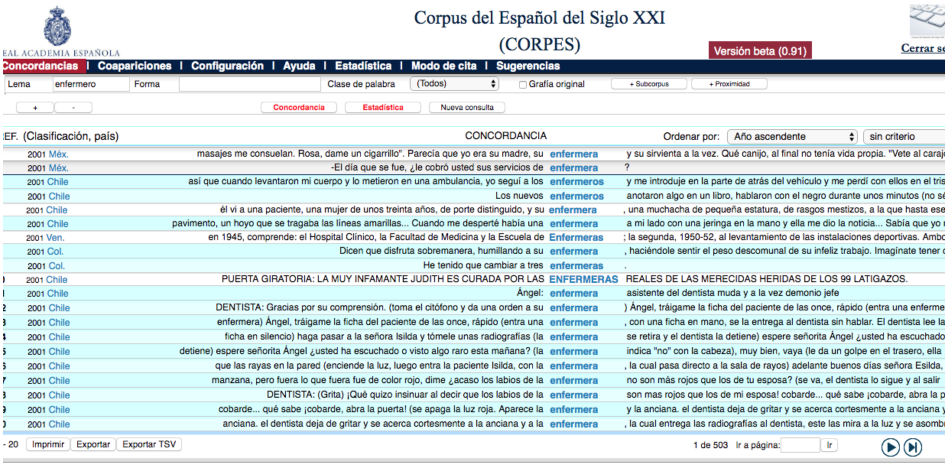


Figure 3: Rate of Frequency of *la enfermera* and *el enfermero* in Google Books Ngrams

Furthermore, searching the Spanish word *enfermera* in the lemma search system of CORPES XXI retrieves no hits. However, searching *enfermero* ('male nurse') in this search system retrieves more than 10,000 hits, most of which are for the word *enfermera*, as shown in the first concordance (Figure 4):



Lema: enfermero
Forma:
Clase de palabra: (Todos)
□ Grafía original

Ordenar por: Año ascendente | sin criterio

EF. (Clasificación, país)	CONCORDANCIA		
2001 Méx.	masajes me consuelan. Rosa, dame un cigarrillo". Parecía que yo era su madre, su	enfermera	y su sirvienta a la vez. Qué canijo, al final no tenía vida propia. "Vete al caraj
2001 Méx.	-El día que se fue, ¿le cobró usted sus servicios de	enfermera	?
2001 Chile	así que cuando levantaron mi cuerpo y lo metieron en una ambulancia, yo seguí a los	enfermeros	y me introduje en la parte de atrás del vehículo y me perdí con ellos en el tris
2001 Chile	Los nuevos	enfermeros	anotaron algo en un libro, hablaron con el negro durante unos minutos (no sé
2001 Chile	él vi a una paciente, una mujer de unos treinta años, de porte distinguido, y su	enfermera	, una muchacha de pequeña estatura, de rasgos mestizos, a la que hasta ese
2001 Chile	pavimento, un hoyo que se tragaba las líneas amarillas... Cuando me desperté había una	enfermera	a mi lado con una jeringa en la mano y ella me dio la noticia... Sabía que yo
2001 Ven.	en 1945, comprende: el Hospital Clínico, la Facultad de Medicina y la Escuela de	Enfermeras	; la segunda, 1950-52, al levantamiento de las instalaciones deportivas. Ambo
2001 Col.	Dicen que disfruta sobremanera, humillando a su	enfermera	, haciéndole sentir el peso descomunal de su infeliz trabajo. Imagínate tener
2001 Col.	He tenido que cambiar a tres	enfermeras	.
2001 Chile	PUERTA GIRATORIA: LA MUY INFAMANTE JUDITH ES CURADA POR LAS	ENFERMERAS	REALES DE LAS MERECIDAS HERIDAS DE LOS 99 LATIGAZOS.
2001 Chile	Ángel:	enfermera	asistente del dentista muda y a la vez demonio jefe
2001 Chile	DENTISTA: Gracias por su comprensión. (toma el citófono y da una orden a su	enfermera	) Ángel, tráigame la ficha del paciente de las once, rápido (entra una enferme
2001 Chile	enfermera) Ángel, tráigame la ficha del paciente de las once, rápido (entra una	enfermera	, con una ficha en mano, se la entrega al dentista sin hablar. El dentista lee la
2001 Chile	ficha en silencio) haga pasar a la señora Isilda y tómele unas radiografías (la	enfermera	se retira y el dentista la detiene) espere señorita Ángel ¿usted ha escuchado
2001 Chile	detiene) espere señorita Ángel ¿usted ha escuchado o visto algo raro esta mañana? (la	enfermera	indica "no" con la cabeza), muy bien, vaya (le da un golpe en el trasero, ella
2001 Chile	que las rayas en la pared (enciende la luz, luego entra la paciente Isilda, con la	enfermera	, la cual pasa directo a la sala de rayos) adelante buenos días señora Esilda,
2001 Chile	manzana, pero fuera lo que fuera fue de color rojo, dime ¿acaso los labios de la	enfermera	no son más rojos que los de tu esposa? (se va, el dentista lo sigue y al salir
2001 Chile	DENTISTA: (Grita) ¡Qué quizo insinuar al decir que los labios de la	enfermera	son mas rojos que los de mi esposa! cobarde... qué sabe ¡cobarde, abra la p
2001 Chile	cobarde... qué sabe ¡cobarde, abra la puerta! (se apaga la luz roja. Aparece la	enfermera	y la anciana. el dentista deja de gritar y se acerca cortesmente a la anciana y
2001 Chile	anciana. el dentista deja de gritar y se acerca cortesmente a la anciana y a la	enfermera	, la cual entrega las radiografías al dentista, este las mira a la luz y se asomb

- 20

1 de 503 Ir a página:

Figure 4: Concordance of *enfermero* in CORPES XXI (first concordance retrieved)

Finally, lexicographers sometimes include female nouns and notes referring to them. The inclusion of headwords such as **jueza**, **médica**, cross-references (e.g. ➔ **fiscal**) and notes such as 'the form *juez* is also used for feminine' reinforce the male dominance seen in general dictionaries of Spanish. Some of the above dictionaries define **jueza** or **médica** as a 'judge or doctor's wife'. This use was common in informal Spanish 40 or 50 years ago. At that time, most people who trained to care for the sick or infirm were female nurses (*enfermera*), and, following the same logic of the headwords **jueza** and **médica**, Spanish lexicographers should have included a definition of *enfermero* as 'a female nurse's husband'. Such a definition was never included, although it was also used in informal spoken Spanish. The only explanation for this is the influence of power and ideology in dictionary making; that is to say, lexicographers assume the ideology of the dominant group (males) and act accordingly (a) by creating lemmas for referring to a man entering a new profession (without doing the same in the case of a woman), or (b) by referring to the marital status of a woman but not of a man.

To combat this, we believe that proscription notes, that is, lexicographic notes recommending uses and meanings, must be included (see the section below).

Briefly, general dictionaries of Spanish are examples of the influence of the power and ideology that reinforce male dominance, contributing to male leadership and maintaining the norms and expectations of the most powerful group, by, for instance, *explaining* that masculine terms are generic and include feminine ones. To finish with such gendered practices, we have created a new type of dictionaries: the *Diccionarios Valladolid-UVa.* This is an integrated dictionary portal. Fuertes-Olivera (2016) defines it as:

> a reference tool whose Dictionary Writing System is equipped with disruptive technologies. These allow lexicographers to store as much data as possible and users to retrieve only the data they need in specific use situations. Its articles are prepared by the same team with the basic aim of helping human and/or machine users to meet their needs in a quick and easy way. They contain both lexicographically prepared data and open linked data with lexicographic value. The lexicographic data is reusable, subject to a constant process of updating and can be used in conjunction with other tools, e.g. assistants.

The above definition shows that our integrated dictionary portal is a *tool,* that is, a utility and information device conceived for consultation with the genuine purpose of meeting users' specific information needs in different extra-lexicographic situations. This concept fully concurs with the idea of lexicography advocated by proponents of the Function Theory of Lexicography (Bergenholtz and Tarp 2003, Fuertes-Olivera and Tarp 2014, Tarp 2008). In practical terms, this means that we envisage an integrated dictionary portal as a repository of lexicographic data dealing with language, facts and things. Hence, this portal aims at offering reference solutions regardless of their being deemed semantic, encyclopaedic, linguistic, onomastic or whatever. At the time of writing this paper, the portal contains around 80,000 definitions of general Spanish headwords, approximately 20,000 definitions of general English headwords and some 15,000 definitions of English and Spanish accounting terms. We plan to publish several online general dictionaries of Spanish, various online bilingual

Spanish–English/English–Spanish dictionaries and several online specialised dictionaries. For this paper, we will refer to decisions concerned with the use of inclusive language in Spanish. For reasons of space, we will limit our analysis to the words *profesora* and *profesor.* These will illustrate the general philosophy of the tool, focusing on offering a de-gendered approach to dictionary making with the aim of eliminating power and ideology from lexicographic practice, offering a fair perspective of women and men, and recording and promoting social changes.

All lexicographic data are extracted from the Internet. We have used crawlers to find out which words Spaniards typically search for, and we are using Google minitexts, that is, the three lines Google shows when searching for a particular word, as sources for understanding and recording the meaning, use, function, etc., of a particular headword (Tarp and Fuertes-Olivera 2016). The data extracted is subjected to different types of analysis, and of relevance here is that concerned with the use of inclusive language in dictionary making. To the best of our knowledge, no scholar has so far addressed this issue in an integrated way. Existing publications only describe the phenomenon of feminisation in dictionaries in current use and do not explain how lexicographers must eliminate gender bias in dictionaries. Baider et al. (2007), for instance, investigate the definitions in entries for the nouns *homme* 'man' and *femme* 'woman' in the online *EuroWordNet* dictionary. Their comparison reveals that "androcentrism still prevails in this online dictionary, since most examples given in the entries refer to males" (Westveer, Sleeman and Aboh 2018: 376).

Similarly, Darmestädter (2011) compares the 8th and 9th editions of the dictionary of the French Academy with the aim of analysing possible differences between both editions concerning "including the feminisation of profession nouns", and observes that the French Academy "still disfavours the use of feminine forms, prescribing the use of compound forms with *femme* (e.g. *femme médecin* 'female doctor') when no feminine form exists" (Westveer et al. 2018: 376).

Epple (2000) investigates diachronic changes in the presence of female-denoting nouns in different editions of bilingual dictionaries covering American English, French, German and Spanish, and finds "considerable progress in the visibility of women among the different editions of the dictionaries with respect to the inclusion of female-denoting nouns". However, as she shows in the examples in the dictionaries' entries of animate nouns, "women are often not included" (Westveer et al. 2018: 377).

Our first decision has been to eliminate headwords such as **professor, ra** from dictionary making. In the *Diccionarios Valladolid-UVa* all headwords are real words, that is, they are presented as they are used. Hence, the 20 words covered in Tables 1 to 5 are lemmatised as **profesora**, **profesor**, **maestro**, **maestra**, **enseñante**, **médico**, **médica**, **enfermero**, **enfermera**, **juez**, **jueza**, **abogado**, **abogada**, **fiscal**, **fiscala**, **funcionario**, **funcionaria**, **director**, **directora** and **auxiliar**.

Secondly, all headwords are grammatically described with their inflexions and the articles with which they typically occur. For instance, the headwords **profesora** and **profesor** are grammatically described as: *una profesora*; *la profesora*; *unas profesoras* and *las profesoras* (In Spanish, these articles are typically associated with feminine nouns); and *un profesor*; *el profesor*; *unos profesores* and *los profesores* (In Spanish, these articles are typically associated with masculine nouns). This allows us to eliminate concepts such as masculine, feminine, common, etc., which are usually used in general dictionaries of Spanish. This decision has two implications: (a) we take a broadly constructionist approach to gender, and view it as being accomplished in interaction rather than as a fixed category; (b) we claim that generics can be constructed with both, say, *las profesoras* and *los profesores.* For instance, for headwords such as **enfermera** our dictionaries record specific and generic definitions, (see below).

Thirdly, the definitions of profession nouns include specific and generic explanations. By way of example, the Spanish word *profesor* is defined as (1) *hombre que se dedica profesionalmente a la enseñanza, es decir, que está especializado en una materia, disciplina académica, ciencia o arte determinadas y enseña a otras personas* ('man who teaches professionally') and (2) *persona (hombre o mujer) que se dedica profesionalmente a la enseñanza, es decir, que está especializada en una materia, disciplina académica, ciencia o arte determinadas y enseña a otras personas* ('person, i.e. man or woman, who teaches professionally'). The generic definition is recorded under the headword that occurs typically in Spanish. For instance, in Google Books Ngram Viewer, we observe that *los profesores* is around 30 times more frequent than *las profesoras* (Figure 5), and therefore we include the generic definition under the headword **profesor**. However, for the word *feminista* ('feminist') the generic meaning is under the headword **feminista** described with the articles *una*, *la*, *unas*, and *las* (Figure 2 shows that *las feministas* is 42 times more frequent than *los feministas*):

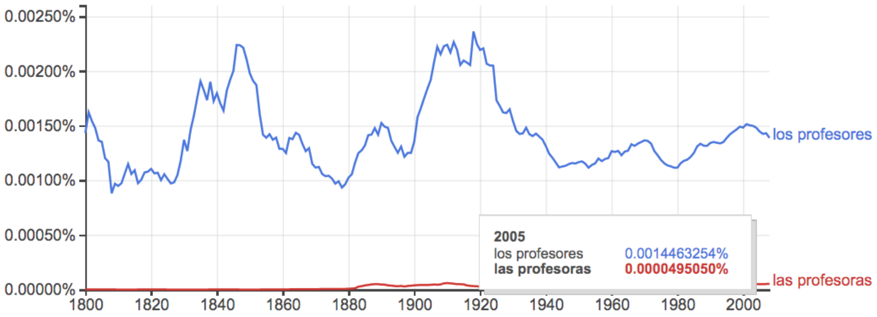


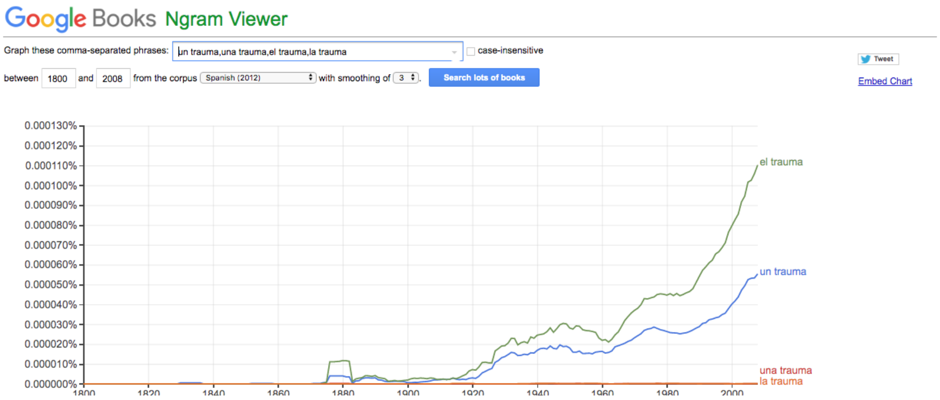
Figure 5: Rate of Frequency of *los profesores* and *las profesoras* in Google Books Ngrams

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1699 (Article)

Fourthly, all definitions use the same style and wording with the exception of the Spanish words *mujer* ('woman'), *hombre* ('man') and *persona* ('person'), which are used at the beginning of the definition of profession nouns for referring, respectively, to a woman, a man, or a person (see the definitions of the word *profesor* above).

Fifthly, each meaning is reinforced with synonyms, antonyms, notes, collocations and examples that are balanced and ideologically neutral. For instance, for the specific meaning of the headword **maestro** we include the synonym *profesor* ('male professor') and for the generic meaning of **maestro** we include the synonyms *profesora* ('female professor') and *profesor* ('male professor'). Then we include notes such as 'this meaning is outdated and should be avoided', referring to the informal meaning of *fiscala*, as 'a male public prosecutor's wife'. Especially relevant is the use of proscription notes. Bergenholtz (2003) has claimed that dictionaries must record the forms and uses found in their sources and, when needed, they should include recommendations. Thus, for the Spanish meaning of the headword **médica** as 'a male doctor's wife' we have included the note *no recomendamos este uso porque hace referencia a un tipo de sociedad que, o bien ya no existe o está desapareciendo por ser discriminatoria para la mujer* (we do not recommend this usage because it refers to a type of society that either does not exist or it is disappearing as it discriminates against women). There are several types of proscription notes, of which those concerned with the use of inclusive language are of relevance for these articles. For instance, for the generic definition of the headword **profesor**, we have also included the synonym *profesores y profesoras* ('male and female professors') and a note indicating that this synonym is typically found in public discourse with the aim of making women visible and eliminating gender bias from Spanish society.

Finally, when we find, say, a male headword whose female counterpart is difficult to spot, for example, because it is a nonce formation, we always search for its female counterpart reproducing typical Spanish patterns of word formation. If, say, we have found the Spanish word *trauma* (informal Spanish for 'male and female orthopaedic surgeon'), then we lemmatise it by attaching suitable inflections and articles. In such situations, we have two headwords, one with the articles *un*, *el*, *unos* and *los*, and another with the articles *un*a, *la*, *unas* and *las*. The former refers to a male orthopaedic surgeon whereas the latter describes a female one. Both headwords are lexicographically treated in the same way. This means that each of them will be described specifically (i.e. referring to a male or female orthopaedic surgeon), and one of them will have a generic meaning (i.e. referring to both male and female orthopaedic surgeons). For deciding which will be defined generically, we basically resort to Google Books Ngram Viewer. For instance, Figure 6 shows that *un trauma* and *el trauma* are much more common than *una trauma* and *la trauma*. Hence, the generic meaning is included in the headword with the articles *un*, *el*, *uno* and *los*:



Graph these comma-separated phrases: un trauma,una trauma,el trauma,la trauma □ case-insensitive

between 1800 and 2008 from the corpus Spanish (2012) with smoothing of 3.

Figure 6: Rate of frequency of *un trauma*, *el trauma*, *una trauma* and *la trauma* in Google Books Ngrams

6. Conclusions

This study has offered a new perspective on language and gender by focusing on gendered practices in general dictionaries of Spanish. We have hypothesised that the lexicographic treatment of very common Spanish words will show to what extent much acclaimed and commonly used dictionaries deal with the question of inclusive language, especially at a time when a Spanish Vice-president has officially requested the elimination of any gender bias, and unfair power and ideology, from official texts and from the *DLE*, the Royal Spanish Academy's dictionary, always regarded as a source of authority. Our hypothesis is based on our assumption that we can observe the possible existence of gendered practices by analysing profession nouns referring to female and male professionals who work in socioeconomic sectors where both are evenly distributed.

Our analysis has focused on two types of function-related texts: headwords and definitions, some of which also include lexicographic notes that are relevant for this study. Regarding our first research question, we have shown that existing general dictionaries of Spanish make women invisible because they use headwords that do not contemplate the existence of women and they use definitions that reproduce the gender bias existing in Spanish speaking societies.

Regarding our second research questions, we have shown that Spanish dictionaries make sometimes women visible, although treating them unfairly, especially by referring to them as 'the wife of a professional man'. This practice does not exist when the new lemma should be coined from an existing female one, e.g. 'modista' or when it could also refer to the husband of a professional woman. In other words, the dictionaries under analysis, which are the most

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1699 (Article)

influential and most used in the Spanish-speaking world, are examples of the influence of the power and ideology that reinforce male dominance, contribute to male leadership and maintain the norms and expectations of the most powerful group. We have made our case that the elimination of such dangerous practices in reference works demands a new approach to dictionary making.

This new approach, in line with our third research question, is based on the concept of the dictionary as both a text and a tool, which must be as balanced as possible and must aim at offering a fair view of society in general and of its members in particular. To achieve this purpose, we have advocated several lexicographic practices, all of which are exemplified with our practice in making the *Diccionarios Valladolid-UVa*. This new type of online dictionaries has eliminated 'dictionarese', has lemmatised profession nouns of women and men on an equal footing, has created balanced and equal definitions, and has introduced new ideas and practices for promoting the use of inclusive language in Spanish. In sum, we

— use real words as lemmas, e.g. 'profesor' and 'profesora' are two lemmas;
— craft identical definitions for 'hombre' (man), 'mujer' (woman) or 'persona' (person), each of which refers to a specific or generic profession noun;
— use proscription notes and other lexicographic devices for emphasizing the use of inclusive language
— eliminate any trace of ideology and power in the lexicographic treatment of all lemmas; and
— base all our lexicographic decisions on data extracted from the internet and analyzed in their sociological contexts.

Acknowledgments

Thanks are due to two anonymous reviewers and to Dr. Elsabé Taljard for their comments on a previous draft of this paper.

References

A. Dictionaries

Clave. 2004. *Diccionario de uso del español actual.* Maldonado, María Concepción (Dir.). Seventh edition. Madrid: SM. Accessed on 22 August 2021. http://clave.smdiccionarios.com/app.php.

DLE. 2014. *Diccionario de la Lengua Española*. Real Academia Española de la Lengua y Asociación de Academias de la Lengua Española. 23rd edition. Madrid: Espasa. Accessed on 22 August 2021. https://dle.rae.es/.

EuroWordNet. Accessed on 22 August 2021. http://projects.illc.uva.nl/EuroWordNet/.

Kramarae, C and Treichler, P. 1985. *A Feminist Dictionary*. London: Pandora.

Lexico by Oxford. Accessed on 22 August 2021. https://www.lexico.com/.

Moliner, María. 2007. *Diccionario de uso del español*. Third edition. Madrid: Gredos.

Seco, Manuel, Olimpia Andrés and Gabino Ramos. 2011. *Diccionario del español actual*. 2nd edition. Madrid: Aguilar.

Sinclair, John (Ed.). 1987. *Collins Cobuild English Language Dictionary*. London: Collins ELT.

B. Other literature

Baider, F., É. Jacquey and A. Liang. 2007. La place du genre dans les bases de données multilingues: le cas d'*EuroWordNet*. *Nouvelles Questions Feministes* 26(3): 57-69.

Baron, D. 1986. *Grammar and Gender*. New Haven: Yale University Press.

Baxter, J. 2010. *The Language of Female Leadership*. Basingstoke: Palgrave Macmillan.

Bergenholtz, H. 2003. User-oriented Understanding of Descriptive, Proscriptive and Prescriptive Lexicography. *Lexikos* 13: 65-80.

Bergenholtz, H. and S. Tarp. 2003. Two Opposing Theories: On H.E. Wiegand's Recent Discovery of Lexicographic Functions. *Hermes, Journal of Linguistics* 31: 171-196.

Bergenholtz, H., S. Tarp and H.E. Wiegand. 1999. Datendistributionsstrukturen, Makro- und Mikrostrukturen in neueren Fachwörterbüchern. Hoffmann, L., H. Kalverkämper, H.E. Wiegand, together with Christian Galinski and Werner Hüllen (Eds.). 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. Berlin: De Gruyter.

Bosque, I. and M.A. Barrios Rodríguez. 2018. Spanish Lexicography in the Internet Era. Fuertes-Olivera, Pedro A. (Ed.). 2018. *The Routledge Handbook of Lexicography*: 636-660. London/New York: Routledge.

Cabeza Pereiro, M.C. and S. Rodríguez Barcia. 2013. Aspectos ideológicos, gramaticales y léxicos del sexismo lingüístico. *Estudios Filológicos* 52: 7-27.

Calero Vaquera, M.L. 2010. Ideología y discurso lingüístico: La Etnografía como subdisciplina de la glotopolítica. *Boletín de Filología* 45(2): 31-48.

Chen, W.G. 2015**.** Bilingual Lexicography and Recontextualization: A Case Study of Illustrative Examples in a New English–Chinese Dictionary. *Australian Journal of Linguistics* 35(4): 311-333.

Chen, W.G. 2019. Towards a Discourse Approach to Critical Lexicography. *International Journal of Lexicography* 32(3): 362-388.

Corpus de Referencia del español actual. Accessed on 22 August 2021. http://corpus.rae.es/creanet.html.

Corpus del Español del Siglo XXI. Accessed on 22 August 2021. http://web.frl.es/CORPES/view/inicioExterno.view.

Darmestädter, C. 2011. Modernité et modernisation du *Dictionnaire de l'Académie française*: quelles transformations de la huitième à la neuvième édition? *Études de linguistique appliquée* 163(3): 285-306.

De Beauvoir, S. 1949. *Le Deuxième Sexe. Les faits et les mythes.* Volume 1. *L'expérience vécue.* Volume 2. Paris: Gallimard.

Dubois, J and C. Dubois. 1971. *Introduction à la lexicographie: le dictionnaire*. Paris: Larousse.

Epple, B. 2000. Sexismus in Wörterbüchern. 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*: 739-754.

Fairclough, N. 1989. *Language and Power*. London: Longman.

Forgas Berdet, E. 1996. Lengua, sociedad y diccionario: La ideología. Forgas Berdet, E. (Ed.). 1996. *Léxico y diccionarios:* 71-90. Tarragona: Universitat Rovira i Virgili.

Frawley, W. 1989. The Dictionary as Text. *International Journal of Lexicography* 2(3): 231-248.

Fuertes Olivera, Pedro A. 1992. *Mujer, lenguaje y sociedad. Los estereotipos de género en inglés y en español.* Madrid: Ayuntamiento de Alcalá de Henares.

Fuertes-Olivera, Pedro A. 2007. A Corpus-based View of Lexical Gender in Written Business English. *English for Specific Purposes* 26(2): 219-234.

Fuertes-Olivera, Pedro A. 2016. *European Lexicography in the Era of the Internet: Present Situations and Future Trends.* Plenary talk, Beijing, 2 December 2016. Talk sponsored by the Commercial Press and the Chinese Association of Lexicography.

Fuertes-Olivera, Pedro A. (Ed.). 2018. *The Routledge Handbook of Lexicography*. London: Routledge.

Fuertes-Olivera, Pedro A. and A. Arribas-Baño. 2008. *Pedagogical Specialised Lexicography. The Representation of Meaning in English and Spanish Business Dictionaries.* Amsterdam/Philadelphia: John Benjamins.

Fuertes-Olivera, Pedro A. and S. Nielsen. 2018. Translating English Specialized Dictionary Articles into Danish and Spanish: Some Reflections. *3L: Language, Linguistics, Literature* 24(3): 15-25.

Fuertes-Olivera, Pedro A. and S. Tarp. 2014. *Theory and Practice of Specialised Online Dictionaries. Lexicography versus Terminography.* Berlin/Boston: De Gruyter.

Google Books Ngram Viewer. Accessed on 3 February 2021. https://books.google.com/ngrams.

Halliday, M.A.K. 1978. *Language as Social Semiotic: The Social Interpretation of Language and Meaning.* London: Arnold.

Hanks, P. 2012. The Corpus Revolution in Lexicography. *International Journal of Lexicography* 25(4): 398-436.

Harris, R. and C. Hutton. 2007. *Definition in Theory and Practice. Language, Lexicography and the Law.* London: Continuum.

Hidalgo Tenorio, E. 2000. Gender, Sex and Stereotyping in the *Collins COBUILD English Language Dictionary*. *Australian Journal of Linguistics* 20(2): 211-230.

Holmes, J. 2005. Power and Discourse at Work: Is Gender Relevant? Lazar, Michelle M. (Ed.). 2005. *Feminist Critical Discourse Analysis*: 31-60. London: Palgrave.

Holmes, J. 2006. *Gendered Talk at Work.* Oxford: Blackwell.

Holmes, J. and B.W. King. 2017. Gender and Sociopragmatics. Barron, Anne, Yueguo Gu and Gerard Steen (Eds.). 2017. *The Routledge Handbook of Pragmatics:* 121-138. London: Routledge.

Holmes, J. and M. Meyerhoff (Eds.). 2003. *The Handbook of Language and Gender.* Oxford: Blackwell.

Informe sobre el buen uso del lenguaje inclusivo en nuestra carta magna. 16 de enero de 2020 (January, 16, 2020). Accessed on 3 February 2021. https://www.rae.es/noticias/el-pleno-de-la-rae-aprueba-el-informe-sobre-el-buen-uso-del-lenguaje-inclusivo-en-nuestra.

Kachru, B.B. 1995. Afterword: Directions and Challenges. Kachru, Braj B. and Henry Kahane (Eds.). 1995. *Cultures, Ideologies and the Dictionary. Studies in Honour of Ladislav Zgusta:* 417-424. Tübingen: Max Niemeyer.

Keating, E. 2009. Power and Pragmatics. *Language and Linguistics Compass* 3(4): 996-1009.

Mills, S. and L. Mullany. 2011. *Language, Gender and Feminism: Theory, Methodology and Practice.* London: Routledge.

Ministerio de Trabajo. 2020. Accessed on 3 June 2020. https://www.ine.es/dyngs/INEbase/es/categoria.htm?c=Estadistica_P&cid=1254735976595.

Mullany, L.J. 2007. *Gendered Discourse in the Professional Workplace*. Basingstoke: Palgrave Macmillan.

Nielsen, S. 2011. Function- and User-related Definitions in Online Dictionaries. Kartasova, F.I. (Ed.). 2011. *Ivanovskaya leksikograficheskaya shkola: traditsii i innovatsii:* 197-219. Ivanovo: Ivanovo State University.

Nielsen, S. 2018. LSP Lexicography and Typology of Specialized Dictionaries. Humbley, John, Gerhard Budin and Christer Laurén (Eds.). 2018. *Languages for Special Purposes: An International Handbook:* 71-95. Berlin/Boston: Mouton de Gruyter.

Nissen, U.K. 1986. Sex and Gender Specifications in Spanish. *Journal of Pragmatics* 10(6): 725-738.

Pechenick, E.A., C.M. Danforth and P.S. Dodds. 2015. Characterizing the Google Books Corpus: Strong Limits to Inferences of Socio-Cultural and Linguistic Evolution. *PLoS ONE* 10(10). https://doi.org/10.1371/journal.pone.0137041.

Pichler, P. 2005. Review of J. Holmes and M. Meyerhoff (Eds.). *The Handbook of Language and Gender. Language in Society* 34(4): 633-638.

Rodríguez Barcia, S. 2012. El análisis ideológico del discurso lexicográfico: una propuesta metodológica aplicada a diccionarios monolingües del español. *Verba. Anuario Galego de Filoloxía* 39: 135-159.

Rundell, M. 2015**.** From Print to Digital: Implications for Dictionary Policy and Lexicographic Conventions. *Lexikos* 25: 301-322.

Simpson, P. and A. Mayr. 2010. *Language and Power: A Resource Book for Students*. London: Routledge.

Tannen, D. 1990. *You Just Don't Understand: Women and Men in Conversation.* New York: Morrow.

Tarp, S. 2008. *Lexicography in the Borderland between Knowledge and Non-knowledge: General Lexicographical Theory with Particular Focus on Learner's Lexicography.* Tübingen: Max Niemeyer.

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.

Van Dijk, T.A. 1984. *Prejudice in Discourse*. Amsterdam: John Benjamins.

Van Dijk, T.A. 1987. *Communicating Racism: Ethnic Prejudice in Thought and Talk.* Sage: Newbury Park, CA.

Velasco-Sacristán, M. and Pedro A. Fuertes-Olivera. 2006. Towards a Critical Cognitive-Pragmatic Approach to Gender Metaphors in Advertising English. *Journal of Pragmatics* 38(11): 1982-2002.

Westveer, T., P. Sleeman and E.O. Aboh. 2018. Discriminating Dictionaries? Feminine Forms of Profession Nouns in Dictionaries of French and German. *International Journal of Lexicography* 31(4): 371-393.

Wodak, R. 1999. Critical Discourse Analysis at the End of the 20th Century. *Research on Language and Social Interaction* 32(1–2): 185-193.

Wodak, R. and M. Meyer. 2015. Critical Discourse Studies: History, Agenda, Theory and Methodology. Wodak, R. and Michael Meyer (Eds.). 2015. *Methods of Critical Discourse Studies:* 18-50. Third edition. London: Sage.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1700 (Article)

The Intellectualization of African Languages through Terminology and Lexicography: Methodological Reflections with Special Reference to Lexicographic Products of the University of KwaZulu-Natal

Langa Khumalo, *South African Centre for Digital Language Resources, North West University, Potchefstroom, South Africa*
(langa.khumalo@nwu.ac.za)
and
Dion Nkomo, *School of Languages and Literatures: African Language Studies, Rhodes University, Makhanda, South Africa*
(d.nkomo@ru.ac.za)

Abstract: Terminology development and practical lexicography are crucial in language intellectualization. In South Africa, the Department of Sport, Arts and Culture, National Lexicography Units, universities, commercial publishers and other organizations have been developing terminology and publishing terminographical/lexicographical resources to facilitate the use of African languages alongside English and Afrikaans in prestigious domains. Theoretical literature in the field of lexicography (e.g., Bergenholtz and Nielsen (2006); Bergenholtz and Tarp (1995; 2010); Gouws 2020) has attempted to resolve traditional distinctions between lexicography and terminology while also addressing terminological imprecisions in the relevant scholarship. Taking the cue from such scholarship, this article reflects on the methodological approaches for developing lexicographical products for specific subject fields, i.e., resources that document and describe terminology from specialized academic and professional fields. Its focus is on the use of traditional methods vis-à-vis the application of electronic corpora and its technologies in the key practical tasks such as term extraction and lemmatization. The article notes that the limited availability of specialized texts in African languages hampers the development and deployment of advanced electronic corpora and its applications to improve the execution of terminological and lexicographical tasks, while also enhancing the quality of the products. The *Illustrated Glossary of Southern African Architectural Terms (English–isiZulu), A Glossary of Law Terms* (English–isiZulu) and the forthcoming isiZulu dictionary of linguistic terms are used for special reference.

Keywords: INTELLECTUALIZATION OF AFRICAN LANGUAGES, LEXICOGRAPHY, TERMINOLOGY, TERMINOGRAPHY, DICTIONARY, SUBJECT FIELD DICTIONARIES, SUBJECT FIELD

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1700 (Article)

LEXICOGRAPHY, GLOSSARY, ELECTRONIC CORPORA

Opsomming: Die intellektualisering van Afrikatale deur middel van die terminologie en leksikografie: Metodologiese gedagtes met spesifieke verwysing na leksikografiese produkte van die Universiteit van KwaZulu-Natal. Terminologieontwikkeling en praktiese leksikografie is noodsaaklik in taalintellektualisering. In Suid-Afrika het die Departement van Sport, Kuns en Kultuur, die Nasionale Lesikografie-eenhede, universiteite, kommersiële uitgewers en ander organisasies die terminologie ontwikkel en terminologiese/leksikografiese hulpbronne gepubliseer om die gebruik van Afrikatale neffens Engels en Afrikaans in toonaangewende domeine te bevorder. Teoretiese literatuur in die leksikografieveld (soos Bergenholtz en Nielsen (2006); Bergenholtz en Tarp (1995; 2010); Gouws 2020) het pogings aangewend om die tradisionele onderskeid tussen die leksikografie en die terminologie te ontleed en terselfdertyd die terminologiese onjuisthede in die relevante studieveld aan te spreek. Vanuit hierdie agtergrond neem dié artikel die metodologiese benaderings tot die ontwikkeling van leksikografiese produkte vir spesifieke onderwerpsvelde, m.a.w. hulpbronne wat die terminologie van gespesialiseerde akademiese en professionele velde dokumenteer en beskryf, in oënskou. Daar word gefokus op die gebruik van tradisionele metodes versus die gebruik van elektroniese korpora en die tegnologie daaraan verbonde in die belangrikste praktiese take soos term-onttrekking en lemmatisering. In die artikel word daarop gewys dat die beperkte beskikbaarheid van gespesialiseerde tekste in Afrikatale die ontwikkeling en benutting van gevorderde elektroniese korpora en die toepassings daarvan verhinder om sodoende die uitvoer van terminologiese en leksikografiese take te verbeter en terselfdertyd die kwaliteit van die produkte te verhoog. Die *Illustrated Glossary of Southern African Architectural Terms (English–isiZulu)*, *A Glossary of Law Terms* (English–isiZulu) en die toekomstige isiZulu woordeboek van linguistiese terme word as spesifieke verwysing gebruik.

Sleutelwoorde: INTELLEKTUALISERING VAN AFRIKATALE, LEKSIKOGRAFIE, TERMINOLOGIE, TERMINOGRAFIE, WOORDEBOEK, SPESIALEVELDWOORDEBOEKE, SPESIALEVELDLEKSIKOGRAFIE, GLOSSARIUM, ELEKTRONIESE KORPORA

1. Introduction

In South Africa, the declaration of nine indigenous languages as official languages, alongside Afrikaans and English, is yet to achieve the envisaged parity of esteem of all the official languages. English continues to dominate prestigious professional and academic spaces at the expense of mother-tongue speakers of other official languages. Government departments have expressed commitment towards multilingualism by formulating and adopting language policies as per the imperatives of the *Use of Official Languages Act*, while institutions of higher learning have done likewise in response to the *Language Policy for Higher Education*. However, the implementation of language policies in ways that promote multilingualism and parity of esteem among the official languages remains elusive. Multilingualism in official government communication, including the translation of important official documents, as well as the use of

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1700 (Article)

African languages as academic languages in the country's universities, remains handicapped by terminological problems. According to Alberts (2017: 148), terminology is thus "a strategic resource and has an important role in the functional development of a country's languages and their users — especially in a multilingual country".

Indeed, the collection, creation, documentation and description of terminology, generally referred to as terminography, remains a vital undertaking for the intellectualization of African languages. In this contribution, we follow the guidance in Bergenholtz and Tarp (1995; 2010) and Bergenholtz and Nielsen (2006) who dismiss the existence of fundamental disciplinary differences between terminology, particularly terminography, and specialized lexicography. While we recognize their flexible approach in favour of specialized lexicography, for this article we embrace further meticulous disambiguation by Gouws (2020), who indicates that *subject field lexicography* is the more precise term for the branch of lexicography concerned with dictionaries that deal with language or knowledge of specialized disciplines, and subsequently subject field dictionaries as the products of this field. In so doing, we are recognizing as dictionaries even the rudimentary products by compilers with various professional disciplinary inclinations, including those who would not recognize themselves as lexicographers. For example, some of the compilers regard themselves as terminologists, translators or just subject specialists who seek to provide cognitive and communicative support to non-experts, e.g., students who are challenged by the language used in specific subject fields. This is common in African languages. Our interest is not really on the products per se, i.e., whether they qualify to be called dictionaries, but on the methodologies that are used to perform critical tasks in the compilation of special field dictionaries regardless of their scope and depth. We focus on the identification of terms from various sources for lemmatization and lexicographical treatment, as well as the preceding activities, bearing in mind the fact that terminology development remains an integral part of compiling special field dictionaries in African languages. We are interested in reflecting on methodological advances in this enterprise in the light of electronic corpora and the relevant corpus query tools which have expedited lexicographic processes against the challenges posed by lagging intellectualization of African languages. The experience of compiling three subject field dictionaries at the University of KwaZulu-Natal is used for special reference.

2. The intellectualization of African languages through terminology and lexicography

The imperative to intellectualize African languages for expanded functional use in all spheres of life is vital against centuries of their prolonged neglect in favour of colonial languages from Africa's early encounters with foreign settlers from Europe. In the context of skewed power relations that associated Europe with progress on the one hand and Africa with primitiveness on the other, languages such as English, French and Portuguese dominated all the formal public domains of life which privileged written languages. Without a strong literary history, African languages were relegated to the domestic lives of their speakers and peripheries of the new socio-economic, cultural and political order. This meant that the languages could not keep abreast with the development of the modern society. Havránek (1932: 32) defines intellectualization of a language as:

> [I]ts adaptation to the goal of making possible precise and rigorous, if necessary abstract, statements, capable of expressing the continuity and complexity of thought, that is, to reinforce the intellectual side of speech. This intellectualization culminates in scientific (theoretical) speech, determined by the attempt to be as precise in expression as possible, to make statements which reflect the rigor of objective (scientific) thinking in which the terms approximate concepts and the sentences approximate logical judgements.

While Havránek's description of language intellectualization beyond doubt indicates the mammoth task of intellectualizing African languages today, it is important to put it into perspective. Writing in the preface of his famous dictionary, Samuel Johnson had this to say about the English language in the late 18th century:

> When I took the first survey of my undertaking, I found our speech copious without order, and energetic without rules: wherever I turned my view, there was perplexity to be disentangled, and confusion to be regulated; choice was to be made out of boundless variety, without any established principle of selection; adulterations were to be detected, without the sufferages of any writers of classical reputation or acknowledged authority (Crystal 2005: 21).

Johnson's impression clearly suggests that English could not be used to make precise, rigorous, abstract statements to express complex thoughts in a logical way at the time of his writing. If we compare this to isiXhosa in the impression of one of the foremost 19th century isiXhosa lexicographers, John W. Appleyard, one would argue that isiXhosa bore some vital qualities of an intellectualized language. Appleyard wrote:

> How came (sic) these people or their ancestors, centuries ago, to express them in this way, and to adopt this system of alliteration. No one can tell; but whatever their language is; and whatever may have been its origin, the [isiXhosa speakers] themselves are not an *intellectually* (original emphasis) childish race. In all grammatical variations of form, [the] language is eminently distinguished by system and regularity. It is … correctly spoken by all classes of the community, which is not the case, perhaps, with any of our European tongues. As a very general, if not invariable rule, [an isiXhosa speaker] will never be heard using an ungrammatical expression (Appleyard 1850: 67-68).

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1700 (Article)

The perspective that is needed is that the assessment of language intellectualization ought to be contextualized. In the precolonial context with a stable African epistemological order, African languages would undoubtedly serve their speakers optimally in all their intellectual activities, which the English language could not do during Johnson's time in England. English was a disorderly language in terms of Johnson in comparison to Greek and Latin, which had hegemonic roles in Europe, and other emerging standard languages such as Italian and French, which were benefiting from the work of the language academies (Nkomo 2018). African languages were found wanting with the advent of a new intellectual order in which "an intellectualized language [w]as one which can be used for educating a person from kindergarten to the university and beyond" (Sibayan 1991: 229). What is unquestionable is Sibayan's general identification of the goal of intellectualization as that of developing the language "for use in the controlling domains of language" (Sibayan 1991: 72). The introduction of a new idea of intellectualism at the onset of European colonization was accompanied by a decentring of African languages, leading Kaschula and Nkomo (2019) to argue that the languages were in fact *de-intellectualized* and what they now need is *re-intellectualization* in the context of the new intellectual order that draws on multiplicity of epistemologies.

While the introduction of print in African languages was a significant milestone of their intellectualization for the modern world, it would not be sufficient since the goals of this partial intellectualization did not transcend the use of the languages for evangelization purposes. It is largely in this respect that Gouws (2007) classifies the earliest dictionaries in African languages as externally-motivated, since the dictionaries were primarily for the use of missionaries and other European settlers who wanted to learn the languages rather than for the empowerment of the native speakers. This would include dictionaries that were produced for use within the education system, such as the *Oxford English–Xhosa Dictionary* that was compiled to address the challenges experienced by second language learners of isiXhosa, most of whom were English-mother tongue speakers (Fischer et al. 1985: v). It is, therefore, not easy to talk about the intellectualization of African languages in a context where the interests of the language speakers were not a priority. This is not meant to disregard, for example, lexicographical and terminological work in African languages during the missionary and apartheid period in South Africa. In fact, we concur with Mahlalela-Thusi and Heugh (2002: 255) that present efforts to intellectualize African languages need to take "cognisance of the huge amount of work that has already been undertaken in the past" because "[t]here could be much value in a thorough analysis of both terminology and materials published in the past as this could speed up the process of producing modern and appropriate" resources. However, when we consider the broad aim of intellectualizing African languages, we note that these efforts were limited in the sense that they did not seek to empower the speakers of African languages to use their languages to their optimal level as intellectual resources. It is in recogni-

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1700 (Article)

tion of this limitation that Mesthrie (2008) argues that while it is necessary to use African languages in higher education, the conditions for their use remain insufficient. More work still needs to be done.

National Lexicography Units (NLUs) were established primarily to "conserve, preserve, research and document the official languages concerned, by compiling *a monolingual explanatory dictionary and such other dictionaries* (authors' emphasis) as may be required to satisfy the needs of the target users of that language" (PanSALB 2000: 26). The compilation of monolingual explanatory dictionaries was already firmly established at the Bureau of the Woordeboek van die Afrikaanse Taal (WAT) and Dictionary Unit for South African English (DSAE) for in Afrikaans and English respectively since 1926 and 1969 (Gouws 2007). The envisaged dictionaries were the so-called storehouse of the words of a language which were expected to raise the profile of each official language, particularly the African languages which lacked strong lexicographic traditions.

However, subject field dictionaries only featured anecdotally in the conceptualization of the NLUs through the add-on clause "and such other dictionaries" in the previous quotation. This add-on clause permits the NLUs to produce a variety of spin-off products including school dictionaries. This does not diminish the role of those other dictionaries as they are "required to satisfy the needs of the target users of that language" (PanSALB 2000: 26). They are essential for all the official languages to be used on parity with English in specialized professional and academic disciplines. As Łukasik 2016: 211) puts it, in educational contexts, subject field dictionaries serve "the most important … pedagogical (didactic) function". In African languages, they do this by providing specialized academic terminology, information about terms and their use, as well as the specialized knowledge embedded in the terms. This indeed makes subject field lexicography critical in the intellectualization of previously marginalized languages.

From an organized language planning perspective, the subject field and terminological needs of speakers of African languages are primarily meant to be served by the Department of Sport, Arts and Culture (DSAC). According to Alberts (2017), through the Terminology Coordination Section, the DSAC was tasked with the responsibility of developing terminology and publishing terminological dictionaries. To that end, DSAC has produced several multilingual terminology lists whose compilers also refer to as dictionaries (http://www.dac.gov.za/terminology-list). These include the following:

— *Multilingual Pharmaceutical Terminology List*

— *Multilingual Financial Terminology List*

— *Multilingual Human, Social, Economic and Management Sciences Terminology List*

— *Multilingual Natural Sciences and Technology Term List (Sesotho)*

— *Multilingual Natural Sciences and Technology Term List (Tshiveṋḓa–Xitsonga)*

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1700 (Article)

- — *Multilingual Natural Sciences and Technology Term List (Nguni)*
- — *Multilingual Mathematics Dictionary: Grade R–6*
- — *Multilingual HIV/Aids Terminology*
- — *Multilingual Parliamentary/Political Terminology*
- — *Multilingual Terminology for Information Communication Technology*

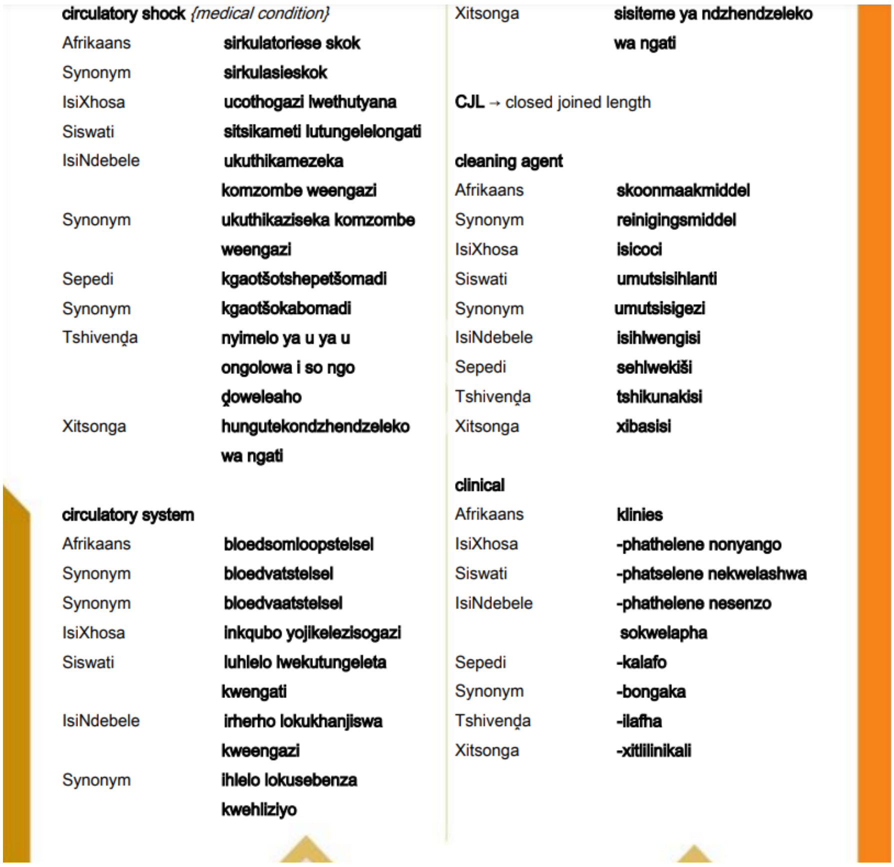
The DSAC has produced most of the above-listed resources under its "Schools Project" which is dedicated to the "documentation of existing terminology, and facilitation of the development of terminology in the African languages for new concepts that appear in the teaching materials for Grades 1 to 6" (DAC 2013a: v). The same motivation has inspired the production of more or less similar products by the Project for the Study of Alternative Education in South Africa (PRAESA), which compiled the *Illustrated Multilingual Science and Technology Dictionary — Intermediate Phase (English–Afrikaans–Xhosa)*. Commercial publishers have also published a few multilingual subject field dictionaries for use within the education system. Examples include the Maskew Miller Longman's *Longman Multilingual Maths Dictionary for South African Schools: English, isiXhosa, Afrikaans* and Cambridge University Press's *Isichazi-magama seziBalo Sezikolo saseCambridge*. The source of the motivation is the *Language-in-Education Policy* (LiEP), adopted in 1997, which acknowledges "the cognitive benefits […] of teaching through one's medium (home language)". A similar motivation derived from the *Language Policy for Higher Education* (LPHE) of 2002 has motivated subject field lexicography that seeks to produce tools that support the use of African languages in higher education. The LPHE expressly identifies dictionaries as necessary for the effective infusion of African languages in higher education. The production of multilingual academic terminology resources (glossaries) is a key activity in South African universities, see in this regard the Open Education Resource Term Bank (OERTB, http://oertb.tlterm.com/), which was a government-funded project, jointly run by the University of Pretoria and the University of Cape Town. The three dictionaries produced at UKZN, which serve as major references in this paper, are further examples.

3. Quality issues of subject field dictionaries in African languages

The production of subject field dictionaries in African languages has been under-researched and under-theorized compared to other dictionary types. However, this is not peculiar to African languages. Gouws (2020: 244) quotes Kilgarriff (2012) who emphasizes that "general language dictionaries are central to the lexicographical firmament", and this includes the space in dictionary research and lexicographic theory. Dictionary criticism has expressed concern with the quality of subject field dictionaries in African languages. According to Gouws (2013: 52), "[…] lack of concern with LSP dictionaries [has] led in far too

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1700 (Article)

many cases to LSP dictionaries not really qualifying as dictionaries but merely playing an inferior role as word lists or other restricted (and often handicapped) reference products". The articles from DASC's *Multilingual Pharmaceutical Terminology List* (http://www.dac.gov.za/sites/default/files/terminology/Multilingual%20Pharmaceutical%20Terminology%20List.pdf) shown below illustrate this concern.



circulatory shock *{medical condition}*	
Afrikaans	**sirkulatoriese skok**
Synonym	**sirkulasieskok**
IsiXhosa	**ucothogazi lwethutyana**
Siswati	**sitsikameti lutungelelongati**
IsiNdebele	**ukuthikamezeka komzombe weengazi**
Synonym	**ukuthikaziseka komzombe weengazi**
Sepedi	**kgaotšotshepetšomadi**
Synonym	**kgaotšokabomadi**
Tshivenḓa	**nyimelo ya u ya u ongolowa i so ngo ḓoweleaho**
Xitsonga	**hungutekondzhendzeleko wa ngati**
circulatory system	
Afrikaans	**bloedsomloopstelsel**
Synonym	**bloedvatstelsel**
Synonym	**bloedvaatstelsel**
IsiXhosa	**inkqubo yojikelezisogazi**
Siswati	**luhlelo lwekutungeleta kwengati**
IsiNdebele	**irherho lokukhanjiswa kweengazi**
Synonym	**ihlelo lokusebenza kwehliziyo**
Xitsonga	**sisiteme ya ndzhendzeleko wa ngati**
CJL → closed joined length	
cleaning agent	
Afrikaans	**skoonmaakmiddel**
Synonym	**reinigingsmiddel**
IsiXhosa	**isicoci**
Siswati	**umutsisihlanti**
Synonym	**umutsisigezi**
IsiNdebele	**isihlwengisi**
Sepedi	**sehlwekiši**
Tshivenḓa	**tshikunakisi**
Xitsonga	**xibasisi**
clinical	
Afrikaans	**klinies**
IsiXhosa	**-phathelene nonyango**
Siswati	**-phatselene nekwelashwa**
IsiNdebele	**-phathelene nesenzo sokwelapha**
Sepedi	**-kalafo**
Synonym	**-bongaka**
Tshivenḓa	**-ilafha**
Xitsonga	**-xitlilinikali**

Figure 1: Articles from the *Multilingual Pharmaceutical Terminology List*

The *Multilingual Pharmaceutical Terminology List* is a typical example of the publications of the DSAC within the Schools Project. While the publications provide the much-needed multilingual terminology to facilitate the use of African languages in education and other areas, the users are not provided with sufficient information that facilitates an understanding and appropriate use of the terms. With most of these products targeted at school learners, they could have been more impactful with additional explanatory and illustrative data.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1700 (Article)

Indeed, most of them are generally rudimentary multilingual terminology lists in which the word *dictionary* is used tentatively in introductory texts but not on the covers.

Quality issues in subject field dictionaries in African languages do not only manifest themselves in the form of limited data. Nkomo (2019) also identifies inclusion of irrelevant data in relation to the target users of some dictionaries, even though this is a less prevalent problem. Examples include part of speech data and tonal marking in dictionaries that will be used in specialized fields where the teaching of grammar is not a priority. In such cases, one notes that compilers of subject field dictionaries merely copy practices and procedures from other dictionary types with different purposes. Ironically, while doing so, the compilers often neglect vital lexicographical aspects such as the planning of dictionary structure. Microstructures and outer texts are underutilized in the planning of subject field dictionaries to enhance the quality of presentation and accessibility of dictionary contents. Gouws (2020) demonstrates that dictionary structure is equally important in subject field dictionaries when he writes:

> Where the compiler of such a dictionary takes the necessary cognizance of guidelines from a general theory of lexicography such a dictionary can become a good dictionary not only on account of the contents but also due to the appropriate dictionary structures and an adherence to the user-perspective and the relevant lexicographic functions (Gouws 2020: 167).

However, the most crucial quality issue with some subject field dictionaries stems from undefined dictionary databases and haphazard lemma section. This is an issue that the subsequent sections of this paper focus on, first demonstrating how term harvesting and description have generally been approached in African languages before focusing on the UKZN projects. We consider this to be a crucial issue because it may result in the exclusion of critical subject terminology that the users need the most in order to use African languages in the high function domains. As crucial tools in the intellectualization of languages, subject field dictionaries in African languages need to be produced in such way that culminates from a scientific language documentation and explication process capable of reflecting the rigor of objective thinking and logical expression.

Nkomo (2019: 104) avers that a major source of quality problems in subject field dictionaries is that "far too often, they are … constructed by everybody". Generally, most of the resources that may be classified as subject field dictionaries in African languages are compiled by subject-field experts without sufficient lexicographic insight, terminologists, translators and even lexicographers who over-rely on subject-field experts. The main motivation is usually terminology development, after which little consideration is given to explanatory and usage data in relation to the terms, as well as the design and presentation issues of the products in which the terms are accessed. While we do not pre-

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1700 (Article)

scribe who should produce subject field dictionaries, given their interdisciplinary nature, the production of subject field dictionaries needs to be collaborative ventures in which there ought to be a great awareness, meticulous and even creative application of lexicographic principles in order to raise the quality of the products for the benefit of the users who need to get optimal information with high levels of user-friendliness. This remains a challenge in African languages and this challenge is closely associated with the methodologies that are currently being used for key compilation processes.

4. Methodological challenges for subject field dictionaries

Although Tarp (2012) draws his examples from Europe to highlight some challenges of specialized lexicography, his characterization of progress made in this field aptly captures the situation in African languages. Tarp (2012) notes that while the two decades preceding the time of his writing witnessed a proliferation of products under this branch of lexicography, such high-level activity and output upsurge are not matched by quality improvement. He attributes what he regards as disappointing progress in specialized lexicography partly to methodological practices that fail to capitalize on the affordances offered by the developments in science and technology. Likewise, this applies to the situation in African languages.

As noted in the previous section, terminology development remains a major priority enterprise in the intellectualization of African languages. In addition to the DSAC, most higher education institutions in South Africa have engaged in bi- or multilingual terminology projects in order to address the perverse "perception that terminology is an intractable obstacle to the use of African languages in high function domains" (Antia and Ianna 2016: 63). The outcome of such investment in the intellectualization of African languages has been the publication of glossaries and special field dictionaries of varying scope and detail. Apart from the problem of duplication of efforts, a standout common feature in the different projects has been the dominance of what Alberts (2017: 179) calls the translation-oriented approach, which she represents in terms of Figure 2 below. This approach is motivated by the fact that African languages have not made a strong footprint in high function domains, resulting in the paucity of specialized texts and terminological gaps in the languages. Thus, the point of departure is usually English terminology lists that are compiled by or with the assistance of subject field experts and the lists are then translated into African languages. The application of this approach is outlined in detail in *Legal Terminology: Criminal Law, Procedure and Evidence*, an ambitious bilingual explanatory English–Afrikaans/Afrikaans–English dictionary of which the aim was to "compile and publish translated versions in all official languages" (Prinsloo, Alberts and Mollema 2015: iii). The isiXhosa edition, *Isigama Sasemthethweni: Umthetho wolwaphulo-mthetho, wenkqubo nobungqina*, was published in 2019.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1700 (Article)

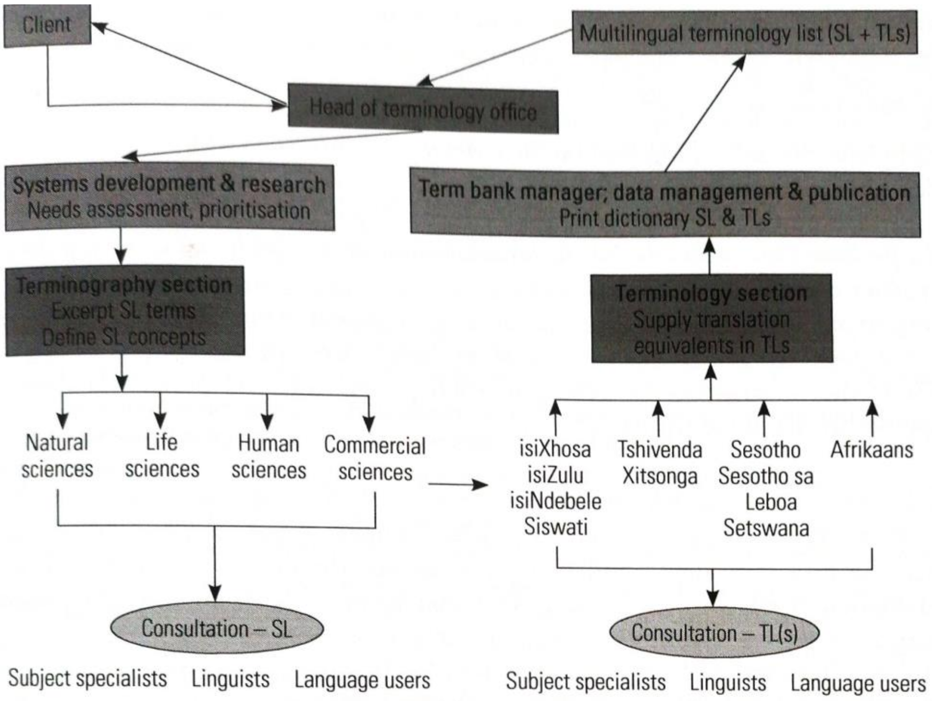


Figure 2: DAC Terminology management process (Alberts 2017: 185)

As illustrated in Fig. 2, in most cases, terminologists and subject experts identify the key concepts that need to be captured and described bi- or multilingually. In the case of university projects linked to specific academic subjects, students are sometimes asked to make submissions of what they have experienced to be challenging concepts for inclusion in the projects. The English terminology lists are usually compiled following a manual term extraction process from relevant sources (Alberts 2017). Unsystematic representation of subject fields may also result from the lack of balance in the selection of English texts, e.g., course outlines and academic textbooks that constitute what would become the dictionary basis from which raw data is drawn for a particular subject field dictionary. Even with a balanced dictionary basis, manual term extraction may result in unbalanced macrostructures with glaring conceptual gaps and incomplete terminological paradigms, as illustrated in Taljard and De Schryver (2002).

In the light of the foregoing, the pioneering exploratory work on corpus applications in African languages lexicography by Danie Prinsloo, Gilles-Maurice de Schryver and Elsabé Taljard, among others, held so much promise in the early 2000s. For example, based on a study on the feasibility of semi-automatic term extraction for the African languages (Taljard and De Schryver 2002: 44), recommended the use of specialized corpora and semi-automatic extraction of terminology in the compilation of subject field dictionaries. They argued that "the semi-automatic extraction of terms for the African languages is not only viable, but even *crucial* in order to counteract inevitable human errors" (Taljard and De Schryver 2002: 66). However, the exciting technological prospects did not blind them to challenges associated with the general level of intellectualization of African languages, as aptly described in the following quote:

> However, if an electronic database is to be compiled for terminological purposes, it presupposes the availability of text material revolving around specific fields. Due to the historically disadvantaged situation of the African languages, even today virtually no subject-specific texts which could be used to build an electronic database are available. As a result of the pre-1994 political and educational system, the vast majority of subject-specific material is written in either English or Afrikaans, with textbooks on literature and grammar of the African languages a possible exception. The African-language terminologist therefore has very little, if any, access to special-field texts which can be used to compile an electronic special-field corpus. This does not only have implications for the compilation of corpora, but also determines the methodology which has hitherto been used by African-language terminologists (Taljard and De Schryver 2002: 47).

While the quotation emphasizes terminology work and terminologists as handicapped by the unavailability of texts in African languages, these problems equally affect translators, lexicographers and virtually all language practitioners who could benefit from specialized corpora. At the time of their writing, the authors were optimistic, though, "that special-language texts will soon be produced on a large scale in the African languages" (Taljard and De Schryver 2002: 47) owing to the official status of the official African languages that was meant to expand their use in the high-status domains. Twenty years on, the situation might have improved, but this would vary according to subject fields, given that English still remains dominant while the use of African languages is regarded as more viable for some subjects, e.g., humanities, than the sciences. This dominance means that African language-texts are mainly produced through translation, which has its own quality challenges as the translations are themselves produced without the assistance of good quality subject field dictionaries and term banks. We are still not in an ideal world where all lexicographic tasks could be automated. In that ideal world, Prinsloo (2014: 1344) compares the role of the lexicographer as that "of the pilot of a fully computerized modern jetliner overseeing processes with limited manual intervention". However, in the real world, Prinsloo (2009: 181) has astutely advised that the corpus "cannot replace the lexicographer, nor should it be regarded as inferior to the knowledge of the lexicographer". The real world of terminology and lexicography in African languages is still dominated by traditional manual processes in which optimal use of specialized electronic corpora still fails to pass the criteria of size, representativeness and balance (Bowker and Pearson 2002). Hence the limited visibility of

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1700 (Article)

corpus applications in the UKZN projects is presented as a major methodological challenge for subject field dictionaries in African languages.

5. The case of subject field dictionaries at UKZN

The intellectualization of isiZulu at UKZN has been driven by the University Language Planning and Development Office (ULPDO) in line with the university's language policy and plan (adopted in 2006 and revised in 2014). The policy seeks to promote the development of isiZulu into an academic language as per national sector imperatives. The development, documentation, description and dissemination of terminology for specialised subject disciplines is at the core of the intellectualization of the isiZulu programme at UKZN and this has culminated in the publication of two works, namely the *Illustrated Glossary of Southern African Architectural Terms* (2016) and *A Glossary of Law Terms* (2018), with an isiZulu dictionary of linguistic terms currently at an advanced stage. This section reflects on the methodological issues in the compilation of special subject field dictionaries in African languages, focusing on the impact of electronic corpora and related technologies.

5.1 Terminology development processes

The University of KwaZulu-Natal designed and adopted a terminology development model that consists of five crucial statutory stages facilitated by the Pan South African Language Board (PanSALB) through its KwaZulu-Natal Provincial office. As captured in Fig. 3, these include:

— harvesting of existing usage terms

— description and translation of terminology that has been harvested or created

— consultation and verification with end-users about the terminology proposed

— authentication and standardization through official national (PanSALB) structures

— "finalization" of the process through the listing of terms on the terminology databases and their publication as reference books for wider institutional and national usage.

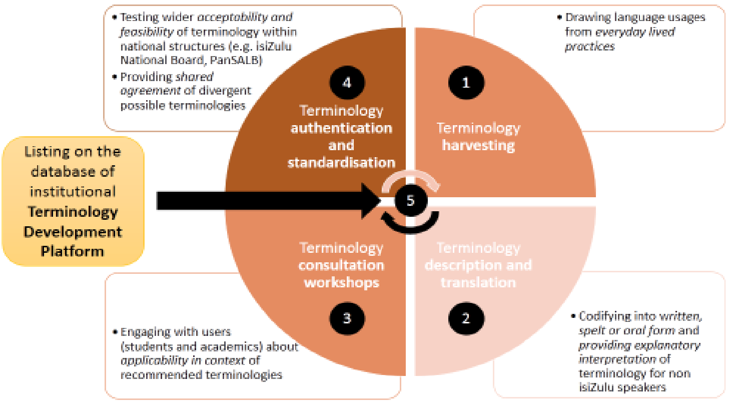

• Testing wider *acceptability and feasibility* of terminology within national structures (e.g. isiZulu National Board, PanSALB)
• Providing *shared agreement* of divergent possible terminologies
• Drawing language usages from *everyday lived practices*
Listing on the database of institutional **Terminology Development Platform**
4
Terminology **authentication and standardisation**
1
Terminology **harvesting**
5
Terminology **consultation workshops**
Terminology **description and translation**
3
2
• Engaging with users (students and academics) about *applicability in context* of recommended terminologies
• Codifying into *written, spelt or oral form* and *providing explanatory interpretation* of terminology for non isiZulu speakers

Figure 3: UKZN terminology development model (Khumalo 2016)

It has been observed in Khumalo (2016) that whereas the language policy at the University of KwaZulu-Natal exists as an important framework for the development of teaching materials in both English and isiZulu, the enforcement of the policy is tepid, cautious and therefore essentially not compulsory. It is in the latter sense that terminology harvesting is done voluntarily by lecturers who are committed to the principles of the language policy, and who also realize the value in making their teaching materials available to students in both languages. The harvested terms are presented as a wordlist of key terms created from a main course/module or a major reference work. It is imperative to state that for the law and architecture dictionaries lemma selection was inspired in part by the critical vocabulary in the discipline as taught at UKZN and the ability by the terminologists and language practitioners on the one hand, and the subject specialists on the other, to successfully find a term equivalent in isiZulu. In the case of the former, the discipline lecturer, who becomes the principal of the discipline terminology development process, would typically lead the process of term harvesting. This would be based on what the lecturer deems as the key English vocabulary that is crucial in the said discipline for the purposes of epistemological access. A standard requirement from the ULPDO is that the initial harvested English term list must not be less than five-hundred words. The English term list must also be accompanied by glosses or definitions that explain the scientific English term and some form of suggested isiZulu equivalent(s) by the discipline lecturer. These are meant to aid the terminologists and the language practitioners in developing and if necessary, coining a cognitively plausible term in isiZulu.

The UKZN terminology development model is largely similar to the approach presented in Figure 2 from Alberts (2017), which is prevalent in multilingual terminology projects in South Africa. In order to broaden the pool

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1700 (Article)

beyond lecturers, crowdsourcing was introduced as a useful strategy to harness discipline specific terminology from multiple individual sources connected to the project. These include lecturers, students, language practitioners, and the general public. The imperative to use crowdsourcing was initiated when ULPDO was developing isiZulu terminology for Information Technology and Computer Science. The two discipline experts, Dr Maria Keet and Dr Graham Barbour created a novel method (cf. http://www.meteck.org/files/commuterm/) of harnessing terms in computer science using computational resources (cf. Keet and Barbour 2014). This proved to be a useful strategy to improve the collection of terminology. It can be observed therefore that the harvesting of terms is a very important exercise as it focuses on the crucial terminology used in the discipline, and is spearheaded by experts, who are informed in the content of the discipline. The terms are then taken through the steps articulated in the model in order to arrive at the isiZulu equivalents, that are made available to the end-users using tools such as the terminology bank and the published pedagogical reference works.

Furthermore, noting the recommendations in studies such as Taljard and De Schryver (2002), the ULPDO has tried to mitigate erratic terminology harvesting, and the effects of a clearly top down and subjective approach to terminology development, by introducing computational applications in an isiZulu dictionary of linguistic terms. This involved the use of the isiZulu National Corpus (INC) of about 1,2 million tokens as a reference corpus as well as an LSP corpus of about 100,000 tokens as a special purpose corpus. The analysis was done using WordSmith Tools, version 6 (https://lexically.net/wordsmith/version6/). It was the objective of the exercise to determine computationally, which words are typical of the linguistic domain in isiZulu and therefore stand out as preferred candidates for headword selection.

The INC as representative of language for general purposes (aka LGP) was used as a reference corpus (RC) and the LSP corpus was used as an analysis corpus (AC). The RC is a non-technical corpus while the AC is a domain-specific, technical corpus. The LSP corpus comprised of the two main isiZulu grammar textbooks *Uhlelo lwesiZulu* and *Izikhali zabaqeqeshi nabafundi*, a collection of isiZulu grammar lecture notes from academics in the School of Arts and the School of Education at UKZN, and some selected online linguistic documents in isiZulu. The aim was to semi-automatically extract terms from the LSP corpus in the subject domain of linguistics. Term extraction remains a challenge to anyone interested in domain-specific information retrieval (Jacquemin 2001; Bourigault et al. 2001). In African languages specifically, the challenges are compounded by the limited availability of specialised texts as the usage of these languages remain restricted in the specialized professional and academic domains.

Table 1 below shows a computationally generated word list (excluding the function words) of linguistic tokens extracted using WS Tools from an LSP corpus. These lemma candidates are generated faster and are presented with corresponding frequency statistical information.

N	Word	Freq.
	a	560
	i	492
	u	375
	e	311
	ubunye	174
	isigaba	167
	o	157
	ubuningi	145
	amagama	120
	amabizo	118
	isenzo	101
	unkamisa	92
	ibizo	78
	inkathi	75
	izenzo	69
	isabizwana	58
	ana	57
	eqondisayo	57
	umoya	57
	ulimi	55
	isivumelwano	46
	izvumelwano	46
	ku	46
	isakhi	41
	iphimbo	39
	imisindo	36
	isandiso	35
	isiqalo	35
	isiqu	35
	izenzukuthi	35
	ukulandula	35
	isibanjalo	34
	edlule	33
	kude	33
	manje	33
	umahluko	33
	isibonelo	31
	sokukhomba	31
	isiqondiso	28
	isahluko	27
	buqamama	26
	soqobo	26
	yesimo	25
	izigaba	24
	oluthambile	23
	yamanje	23
	isikhathi	22
	ukwakhiwa	21
	umsindo	21
	ulwanga	20
	umgudu	20
	isimo	19
	phansi	19
	ungwaqa	17
	phezulu	15
	ubumnini	15
	isichasiso	14
	isijobelelo	14
	uhlelo	13
	amabizoqho	12
	ilunga	12
	isiphawulo	12
	usobizo	12
	isenzukuthi	11
	isizulu	11
	izingasenzo	11
	umenziwa	11
	impambosi	10
	isibaluli	10
	isilandiso	10
	olwangeni	9
	ukulwangisa	9
	umankankana	9
	amabizonto	9
	izizoqho	8
	ikhala	8
	ingqondo	8
	isikhanyiso	8
	isinciphiso	8
	isitho	8
	isingasenzo	7
	umenzi	7
	ichashaza	6
	ifarinksi	6
	imorpheme	6
	ongwaqabathwa	6
	inguqulelo	5
	inhloko	5
	ucezu	5
	umsuka	5
	undebembili	5
	inani	4
	isihlanganiso	4
	isikhuliso	4
	ubulili	4
	ibizomuntu	3
	isibabazo	3
	ugovane	3
	ulwangeni	3
	ukulumbana	3

Table 1: Most frequent 100 linguistic tokens excluding function words

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1700 (Article)

Having created two types of corpora, one a general corpus (the INC) and the other an LSP corpus, it was possible to do a keyness analysis using the keyness function of WS Tools. Table 2 below shows the top 10 tokens in a list of 100 keywords after the keyness analysis.

N	Keyword	English gloss	Freq.	Keyness
1	Isibonelo	example	387	1515,82
2	I	vowel *i*	1002	1424,26
3	A	vowel *a*	1005	1172,94
4	Bese	and	512	875,18
5	Ulimi	language	290	773,57
6	Uma	if	1179	659,00
9	Unkamisa	vowel	180	557,61
10	Mpela	indeed, truly	255	510,56

Table 2: Top 10 of the 100 linguistic tokens

The table shows a typical list of term candidates in the linguistics domain. The keyness tool has successfully extracted candidate terms which are key to the domain of linguistics from the corpus. The list includes the vowels *a, e, i, o, u,* **(3, 11, 2, 38, 13)**; language *ulimi* **(5)**; vowel *unkamisa* **(9)**; singular *ubunye* **(14)**, in a sentence *emshweni* **(15)**; noun class *isigaba* **(16)**, voiceless *ongenazwi* **(18)**; noun *ibizo* **(19)** nouns *amabizo* **(20)**; consonants *ongwaqa* **(39)**; indicative mood *eqondisayo* **(53)**; agreements *izivumelwano* **(59)**; copulative *isibanjalo* **(63)** click sound *ungwaqabathwa* **(68)**; cavity *umgudu* **(80)**; tone *iphimbo* **(87)**; subjectival *senhloko* **(96)**; **etc.**

It was therefore evinced from this extraction process that using such a computationally aided statistical approach is faster, reliable and free from human error or bias. It was again clear that term extraction reduces the amount of noise in the list of candidate terms. However, it can be argued that mother-tongue speaker intuition remains important in complementing this vital computational method (Prinsloo 2009). Human intervention could assist in the inclusion of terms representing conceptual paradigms such as subordination, superordination and coordination relationships. For example, it is possible that the keyness search may provide 'subject concord' as a term but miss out on 'object concord'. The subject field expert can then fill in such a knowledge gap by including such a missing term.

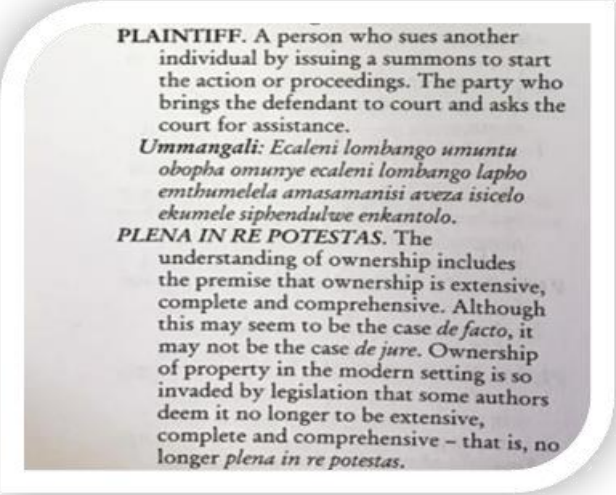
5.2 Some comments on the metadata

The publication of works such as the *Illustrated Glossary of Southern African Architectural Terms* (2016) and the second *A Glossary of Law Terms* (2018) completes stage 5 of the UKZN terminology model and is a culmination of an organic process, which is part of the many terminology dissemination strategies. As noted earlier, the main objective in the whole terminology development process, commencing from the term harvesting of key vocabulary in the discipline by the discipline lecturer, is premised on aiding epistemological access to the subject matter. The terminologists and language practitioners are involved in a process to develop terms that are cognitively plausible and have the potential to improve the understanding of the science in question in the target language. The final product of this terminology development process is therefore aimed to be pedagogical. The two terminology dictionaries are part of the pedagogical tools aimed at improving epistemic access and help improve student success.

While the terminological processes discussed above were rigorous towards the development of scientific terms in isiZulu, the presentation of metadata in the two dictionaries appears to have lacked sufficient theoretical guidance from metalexicography. This has the effect of compromising the quality and utility value of the products. The metadata is sparse and the presentation is characteristically sketchy. Examples in Figure 4 are excerpts from *A Glossary of Law Terms* (2018).

In the case of **PLAINTIF** the headword is presented in capital bold format. The definitions are not numbered. The isiZulu equivalent headword ***Ummangali*** is presented in bold italics. There is no grammatical information. The definition is presented in italics with no usage example. The same treatment is observed with respect to the treatment of **LAW OF CRIMINAL PROCEDURE.** The isiZulu equivalent ***Inqubomthetho yamacala obugebengu/ obulelesi/egazi*** presents a confusing picture. In the absence of a front matter that discusses decisions that are taken in lemma selection and presentation, it is not clear to the user what the slashes stand for and how they relate to the words that come after them. Are they variants of the headword? Are they variants of the last word (as would seem to be the case in this particular lemma)? Would there have been a better way of presenting such information?

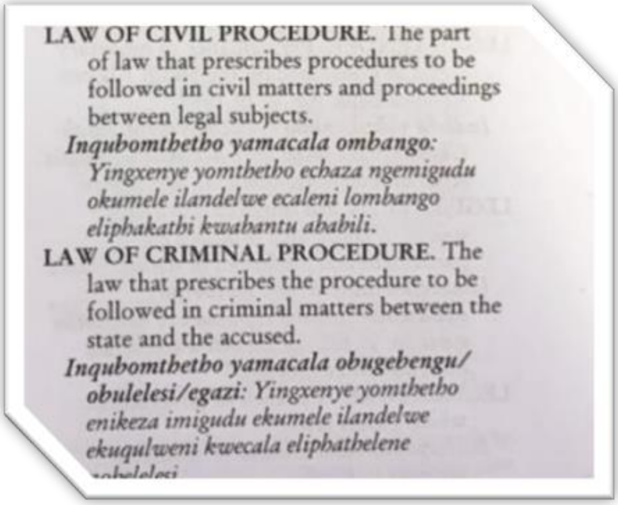
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1700 (Article)



PLAINTIFF. A person who sues another individual by issuing a summons to start the action or proceedings. The party who brings the defendant to court and asks the court for assistance.

Ummangali: *Ecaleni lombango umuntu obopha omunye ecaleni lombango lapho emthumelela amasamanisi aveza isicelo ekumele siphendulwe enkantolo.*

PLENA IN RE POTESTAS. The understanding of ownership includes the premise that ownership is extensive, complete and comprehensive. Although this may seem to be the case *de facto*, it may not be the case *de jure*. Ownership of property in the modern setting is so invaded by legislation that some authors deem it no longer to be extensive, complete and comprehensive – that is, no longer *plena in re potestas.*



LAW OF CIVIL PROCEDURE. The part of law that prescribes procedures to be followed in civil matters and proceedings between legal subjects.

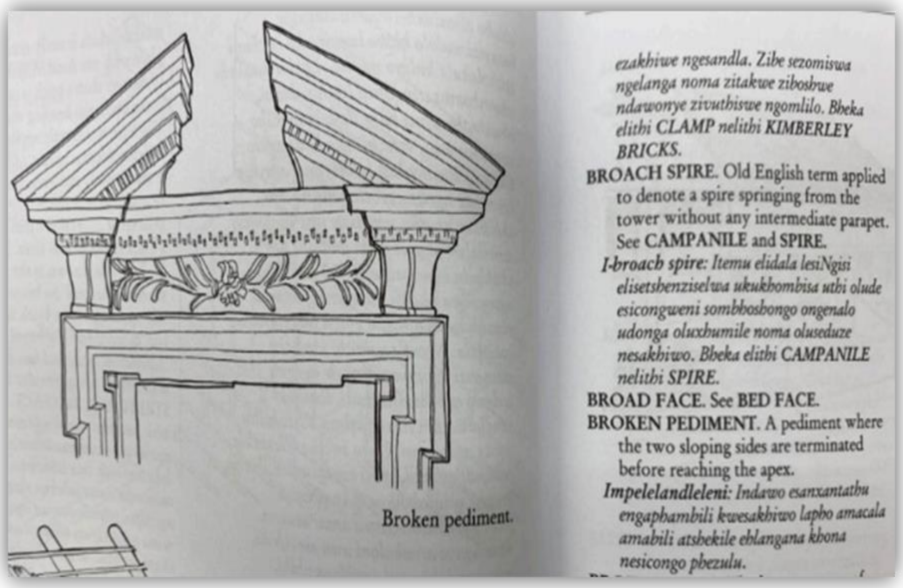
Inqubomthetho yamacala ombango: *Yingxenye yomthetho echaza ngemigudu okumele ilandelwe ecaleni lombango eliphakathi kwabantu ababili.*

LAW OF CRIMINAL PROCEDURE. The law that prescribes the procedure to be followed in criminal matters between the state and the accused.

Inqubomthetho yamacala obugebengu/ obulelesi/egazi: *Yingxenye yomthetho enikeza imigudu ekumele ilandelwe ekuqulweni kwecala eliphathelene yobulelesi*

Figure 4: Examples from the law glossary

Figure 5 below is an example from the *Illustrated Glossary of Southern African Architectural Terms* (2016). While the presentation of the lexicographic material is the same as discussed above, this dictionary has an impressive presentation of illustrations that are key in the discipline of architecture.



Broken pediment.

ezakhiwe ngesandla. Zibe sezomiswa ngelanga noma zitakwe ziboshwe ndawonye zivuthiswe ngomlilo. Bheka elithi CLAMP *nelithi* KIMBERLEY BRICKS.

BROACH SPIRE. Old English term applied to denote a spire springing from the tower without any intermediate parapet. See **CAMPANILE** and **SPIRE.**

I-broach spire: *Itemu elidala lesiNgisi elisetshenziselwa ukukhombisa uthi olude esicongweni sombhoshongo ongenalo udonga oluxhumile noma oluseduze nesakhiwo. Bheka elithi* CAMPANILE *nelithi* SPIRE.

BROAD FACE. See **BED FACE.**

BROKEN PEDIMENT. A pediment where the two sloping sides are terminated before reaching the apex.

Impelelandleleni: *Indawo esanxantathu engaphambili kwesakhiwo lapho amacala amabili atshekile ehlangana khona nesicongo phezulu.*

Figure 5: The *Illustrated Glossary of Southern African Architectural Terms*

Figure 5 presents the article for the lemma **BROKEN PEDIMENT** with isiZulu equivalent ***Impelelandleleni*** and an illustrative drawing of the broken pediment. The inclusion of the illustrations in the *Illustrated Glossary of Southern African Architectural Terms* was an important pedagogical consideration. However, the illustrations are only labelled in English. It is not clear whether this is a lexicographic decision or an omission on the part of the editors as there is no front matter to explain such methodological procedures.

What may be observed is that the compilation of the *Illustrated Glossary of Southern African Architectural Terms* (2016) and *A Glossary of Law Terms* (2018) used the traditional approach. Lemma selection and defining tasks were driven by the subject-field specialists. There was no recourse to an LSP corpus through the use of concordances in order to clarify or illuminate difficult terms. This naturally affected the metadata and influenced the quality of these two terminology dictionaries. Not much consideration was given to issues of dictionary structure by the subject specialists who had neither lexicographic experience nor exposure to lexicographic principles. For instance, the subject-field specialists for the *Illustrated Glossary of Southern African Architectural Terms* (2016) state in the introduction that:

> The idea of publishing this research arose in about 1986, during the course of lectures at the University of Port Elizabeth (*now Nelson Mandela University*) […]. The resultant publication (Frescura 1987) listed about 400 entries written in Eng-

> lish, and brought together for the first time the terminology used by most of the country's language groups, with a primary focus on their historical and rural built environments. Since that time, the original manuscript has undergone extensive additions and revisions, as new research has been undertaken and additional data has become available (Frescura and Myeza 2016: xiv).

The fact that these dictionaries were built within the scope of these existing projects meant that there was very little flexibility in terms of applying the lexicographic theory that the ULPDO staff possessed, besides just converting the presentation of these data sets into a dictionary format.

The compilation of the isiZulu linguistic terms dictionary is a move away from the traditional approach. The publication of the grammar books and other teaching materials in isiZulu means that there was sufficient data to create an LSP corpus. The existence of an LSP corpus also meant that lemma selection could be done using computational approaches through the use of corpus query tools such as WS Tools. Furthermore, the existence of a bigger, IsiZulu National Corpus (the INC), meant that a lot of noise in the lemma selection could be reduced using the keyness approach as explained and demonstrated above. Defining and sense selection has also profited from the use of the concordances when the lemmas are defined. The understanding of lemma concepts does not solely depend on the subject-field specialists, but on the corpus resource as well.

The linguistic terms dictionary is intended to be printed as an A5 medium-sized pocket dictionary, that is portable and user-friendly. Currently in database form, it has just below 5 000 headwords. Size is crucially important for a reference work that is most likely to be in constant use by linguistics students. The dictionary presents lemmas in isiZulu, written in bold lowercase roman letters, followed by the IPA transcription between slashes, followed by tone marking and then the word class, the definition, usage example (optional) and finally its English equivalent. The grammatical information is important since it is part of the familiar jargon in the discipline and is useful for target user comprehension of the discipline. It is notable that such grammatical information might not be as useful in a specialized dictionary of anatomy for instance. Examples below illustrate this point.

uhlelo /úɬɛ|o/ KKP bz 11. DEFINITION. FAN grammar
ibizo /ɪβɪzo/ KKP bz 5. DEFINITION. FAN noun

In addition to the above, the dictionary will have a front matter which provides a brief overview of linguistics as a discipline and a user guide. The lexicographic considerations that have been made in the conceptualization of the isiZulu linguistics terms dictionary make it a potentially more user-friendly resource compared to the other two dictionaries.

6. Conclusion

The development of terminology is an important precursor to the compilation of subject field dictionaries in African languages. The imperative to develop terminology for African languages in South Africa is driven by critical factors that include the repositioning of African indigenous languages in knowledge organization, knowledge creation, knowledge access and knowledge dissemination in (higher) education in order to improve epistemic access and student success, which hitherto has been the bane of higher education. Innovative methodologies are needed in the development, documentation, description and dissemination of terminology, taking advantage of modern advances in technology. While electronic corpus applications have great potential in that respect, as demonstrated in Taljard and De Schryver (2002), limited availability of specialized texts in African languages remains a major hinderance. This means that the benefits of specialized corpora enjoyed by lexicographers, terminologists and translators working on more advanced languages remain a pipedream for those working on African languages. While the article demonstrated that it was possible to maximize on the benefits of electronic corpora in the development of the forthcoming isiZulu dictionary of linguistic terms, it also demonstrated that largely traditional approaches were used in the compilation of the *Illustrated Glossary of Southern African Architectural Terms* and *A Glossary of Law Terms* in isiZulu. These methodological factors had implications on the quality of the products.

Acknowledgement

This article is partly produced within the research programme of the NRF SARChI Chair: Intellectualisation of African Languages, Multilingualism and Education (Grant specific unique reference number 82767), held by the second author. Both authors acknowledge that opinions, findings and conclusions or recommendations expressed in any publication generated by the NRF-supported research are those of the authors, and that the NRF accepts no liability whatsoever in this regard.

References

Alberts, M. 2017. *Terminology and Terminography Principles and Practice: A South African Perspective.* Milnerton: McGillivray Linnegar.

Antia, B. and B. Ianna. 2016. Theorising Terminology Development: Frames from Language Acquisition and the Philosophy of Science. *Language Matters* 47(1): 61-83.

Appleyard, J.W. 1850. *The Kafir Language*. London: Wentworth Press.

Bergenholtz, H. and S. Nielsen. 2006. Subject-field Components as Integrated Parts of LSP Dictionaries. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication* 12(2): 281-303.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1700 (Article)

Bergenholtz, H. and S. Tarp (Eds.). 1995. *Manual of Specialised Lexicography. The Preparation of Specialised Dictionaries*. Amsterdam/Philadelphia: John Benjamins.

Bergenholtz, H. and S. Tarp. 2010. LSP Lexicography or Terminography? The Lexicographer's Point of View. Fuertes-Olivera, P.A. (Ed.). 2010. *Specialised Dictionaries for Learners:* 27-37. Berlin/New York: Walter de Gruyter.

Bourigault, D. et al. 2001. *Recent Advances in Computational Terminology*. Amsterdam/Philadelphia: John Benjamins.

Bowker, L. and J. Pearson. 2002. *Working with Specialized Language: A Practical Guide to Using Corpora.* London: Routledge.

Crystal, D. 2005. *A Dictionary of the English Language: An Anthology.* London: Penguin.

Department of Arts and Culture. 2013a. *Multilingual Mathematics Dictionary: Grade R–6*. Pretoria: Department of Arts and Culture.

Department of Arts and Culture. 2013b. *Multilingual Financial Terminology List*. Pretoria: Department of Arts and Culture.

Department of Arts and Culture. 2013c. *Multilingual HIV/Aids Terminology*. Pretoria: Department of Arts and Culture.

Department of Arts and Culture. 2013d. *Multilingual Parliamentary/Political Terminology*. Pretoria: Department of Arts and Culture.

Department of Arts and Culture. 2013e. *Multilingual Terminology for Information Communication Technology*. Pretoria: Department of Arts and Culture.

Department of Arts and Culture. 2013f. *Multilingual Human, Social, Economic and Management Sciences Terminology List*. Pretoria: Department of Arts and Culture.

Department of Arts and Culture. 2013g. *Multilingual Natural Sciences and Technology Term List (Nguni).* Pretoria: Department of Arts and Culture.

Department of Arts and Culture. 2013h. *Multilingual Natural Sciences and Technology Term List (SeSotho).* Pretoria: Department of Arts and Culture.

Department of Arts and Culture. 2013i. *Multilingual Natural Sciences and Technology Term List (Tshivenḓa–Xitsonga).* Pretoria: Department of Arts and Culture.

Department of Arts and Culture. 2014. *Use of Official Languages Act*. Pretoria: Department of Arts and Culture (DAC).

Department of Arts and Culture. 2021. *Multilingual Pharmaceutical Terminology List.* Pretoria: Department of Arts and Culture (DAC).

Department of Education. 1997. *Language in Education Policy*. Pretoria: Department of Education.

Department of Higher Education and Training. 2002. *Language Policy for Higher Education.* Pretoria: Department of Higher Education and Training.

Deyi, S., G. Minshall and T. Tokwe. 2008. *Longman Multilingual Maths Dictionary for South African Schools: English, isiXhosa, Afrikaans.* Cape Town: Maskew Miller Longman.

Fischer, A., E. Weiss, E. Mdala and S. Tshabe. 1985. *Oxford English–Xhosa Dictionary*. Cape Town: Oxford University Press Southern Africa.

Frescura, F. and J. Myeza. 2016. *Illustrated Glossary of Southern African Architectural Terms: English–isiZulu.* Durban: University of KwaZulu-Natal Press.

Gouws, R.H. 2007. On the Development of Bilingual Dictionaries in South Africa: Aspects of Dictionary Culture and Government Policy. *International Journal of Lexicography* 20(3): 313-327.

Gouws, R.H. 2013. Establishing and Developing a Dictionary Culture for Specialised Lexicography. Jesenšek, V. (Ed.). 2013. *Specialised Lexicography. Print and Digital, Specialised Dictionaries, Databases:* 51-62. Lexicographica Series Maior 144. Berlin/Boston: Walter de Gruyter.

Gouws, R.H. 2020. Special Field and Subject Field Lexicography Contributing to Lexicography. *Lexikos* 30: 143-170.

Havránek, B. 1932. The Functions of Literary Language and its Cultivation. Hávranek, B. and M. Weingart (Eds.). 1932. *A Prague School Reader on Esthetics, Literary Structure and Style:* 32-84. Prague: Melantrich.

Jacquemin, C. 2001. *Spotting and Discovering Terms through Natural Language Processing.* Cambridge, MA: MIT Press.

Kaschula, R.H. and D. Nkomo. 2019. Intellectualisation of African Languages: Past, Present and Future. Wolff, H.E. (Ed.). 2019. *The Cambridge Handbook of African Linguistics*: 601-622. Cambridge: Cambridge University Press.

Keet, M. and G. Barbour. 2014. *Commuterm.* Available at: http://www.meteck.org/files/commuterm.

Khumalo, L. 2016. Disrupting Language Hegemony: Intellectualizing African Languages. Samuel, M., R. Dhunpath and N. Amin (Eds.). 2016. *Disrupting Higher Education Curriculum: Undoing Cognitive Damage*: 247-263. Rotterdam: Sense Publishers.

Kilgarriff, A. 2012. Review of Pedro A. Fuertes-Olivera and Henning Bergenholtz (Eds.). *e-Lexicography: The Internet, Digital Initiatives and Lexicography*. *Kernerman Dictionary News* 20: 26-29.

Łukasik, M. 2016. Specialized Pedagogical Lexicography: A Work in Progress. *Polilog: Studia Neofilologiczne* 6: 211-226.

Mahlalela-Thusi, B. and K. Heugh. 2002. Unravelling some of the Historical Threads of Mother-tongue Development and Use during the First Period of Bantu Education (1955–1975): New Developments and Research. *Perspectives in Education* 20(1): 241-257.

Mbude-Shale, N., Z. Wababa and K. Welman. 2008. *Illustrated Multilingual Science and Technology Dictionary / Isichazi-magama sezeNzululwazi neTeknoloji Ngeelwimi Ezininzi.* Cape Town: New Africa Books.

Mesthrie, R. 2008. Necessary versus Sufficient Conditions for Using New Languages in South African Higher Education: A Linguistic Appraisal. *Journal of Multilingual and Multicultural Development* 29(4): 325-340.

Nkomo, D. 2018. Dictionaries and Language Policy. Fuertes-Olivera, P.A. (Ed.). 2018. *The Routledge Handbook of Lexicography*: 152-165. New York: Routledge.

Nkomo, D. 2019. Theoretical and Practical Reflections on Specialized Lexicography in African Languages. *Lexikos* 29: 96-124.

Nkosi, N.R. and G.N. Msomi. 1992. *Izikhali zabaqeqeshi nabafundi*. Pietermaritzburg: Reachout Publishers.

Nyembezi, S. 1982. *Uhlelo lwesiZulu*. Fourth edition. Pietermaritzburg: Shuter and Shooter.

Open Education Resource Term Bank (OERTB). Available at: http://oertb.tlterm.com/.

PanSALB. 2000. *Annual Report.* Pretoria: Pan South African Language Board (PanSALB).

PRAESA. 2008. *Illustrated Multilingual Science and Technology Dictionary — Intermediate Phase (English–Afrikaans–Xhosa).* Cape Town: New Africa Education.

Prinsloo, D.J. 2009. The Role of Corpora in Future Dictionaries. Nielsen, S. and S. Tarp. (Eds.). 2009. *Lexicography in the 21st Century: In Honour of Henning Bergenholtz*: 181-206. Amsterdam/Philadelphia: John Benjamins.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1700 (Article)

Prinsloo, D.J. 2014. The Utilization of Bilingual Corpora for the Creation of Bilingual Dictionaries. Gouws, R.H., U. Heid, W. Schweickard and H. Wiegand (Eds.). 2014. *Dictionaries. An International Encyclopedia of Lexicography. Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography*: 1344-1356. Berlin/Boston: De Gruyter Mouton.

Prinsloo, M.W., M. Alberts and N. Mollema. 2015. *Legal Terminology: Criminal Law, Procedure and Evidence / Regsterminologie: Straf-, Strafproses- en Bewysreg.* Cape Town: Juta.

Scott, M. 2007. *WordSmith Tools* version 6. Available at https://lexically.net/wordsmith/version6/.

Sibayan, B.P. 1991. The Intellectualisation of Filipino. *International Journal of the Sociology of Language* 88: 69-82.

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.

Tarp, S. 2012. Specialised Lexicography: 20 Years in Slow Motion. *Ibérica. Journal of the European Association of Languages for Specific Purposes* 24: 117-128.

Wababa, Z., K. Welman and K. Press (Eds.). 2010. *Isichazi-magama seziBalo Sezikolo saseCambridge*. Cape Town: Cambridge University Press.

Zondi, K. 2018. *A Glossary of Law Terms: English–isiZulu*. Durban: University of KwaZulu Natal Press.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1701 (Article)

English–Georgian Parallel Corpus and Its Application in Georgian Lexicography

Tinatin Margalitadze, *Centre for Lexicography and Language Technologies, Ilia State University, Georgia (tinatin.margalitadze@iliauni.edu.ge)*

George Meladze, *Centre for Lexicography and Language Technologies, Ilia State University, Georgia (giorgi.meladze.4@iliauni.edu.ge)*
and
Zakharia Pourtskhvanidze, *Institute of Empirical Linguistics, University of Frankfurt, Germany (pourtskhvanidze@em.uni-frankfurt.de)*

Abstract: The Georgian language, the official language of Georgia, is the only written member of the Kartvelian language family, the indigenous language family of the Caucasus region. Georgian philology and lexicography have long-standing tradition, English–Georgian lexicography being no exception.

Given the increasing use of ample electronic text corpora for lexicographical purposes, the team of Georgian lexicographers, working on the *Comprehensive English–Georgian Dictionary* (CEGD), subsequently the *Comprehensive English–Georgian Online Dictionary* (CEGOD), decided to compile an English–Georgian Parallel Corpus (EGPC). The aim of the project was to develop the methodology of building a parallel corpus of Georgian and assess its efficiency for Georgian bilingual lexicography. The work on the corpus is going on for over a decade. The ultimate aim is to create a standard for Georgian bilingual corpora that will be compiled in future.

The article describes the content and composition of the EGPC, its structure, functionalities, search engines and so on. The article also deals with various studies conducted over years in order to assess and enhance the value, applicability and efficiency of the EGPC for the automatic or semi-automatic recognition, tagging and extraction of terminology, the compilation of terminological entries, as well as the entries for the *English–Georgian Dictionary* and those for the *Georgian–English Learner's Dictionary*, etc.

Particular emphasis is laid upon the actual or potential applicability of the corpus for the lexicographical activities and for the machine translation projects. The findings of the study may be interesting for other under-resourced languages like Georgian.

Keywords: PARALLEL CORPUS, TERMINOLOGICAL ENTRIES, ENGLISH–GEORGIAN DICTIONARY, GEORGIAN–ENGLISH DICTIONARY

Opsomming: Die Engels–Georgiese parallelle korpus en die toepassing daarvan in die Georgiese leksikografie. Georgies, die amptelike taal van Georgië, is die enigste geskrewe lid van die Kartveliaanse taalfamilie, die inheemse taalfamilie van die Kaukasiese streek. Die Georgiese taalwetenskap en leksikografie het 'n lang verbintenis waarvan die Engels-Georgiese leksikografie geen uitsondering is nie.

In die lig van die toenemende gebruik van uitgebreide elektroniese tekskorpora vir leksikografiese doeleindes, het die Georgiese span leksikograwe wat aan die *Comprehensive English–Georgian Dictionary* (CEGD), later die *Comprehensive English–Georgian Online Dictionary* (CEGOD), werk, besluit om 'n Engels-Georgiese Parallelle Korpus (EGPK) saam te stel. Die doel van die projek was die ontwikkeling van die metodologie vir die bou van 'n parallelle Georgiese korpus en die bepaling van die effektiwiteit daarvan vir die Georgiese tweetalige leksikografie. Daar word al meer as 'n dekade aan die korpus gewerk. Die uiteindelike doel is om 'n standaard vir Georgiese tweetalige korpora wat in die toekoms saamgestel sal word, te skep.

Die artikel beskryf die inhoud en samestelling van die EGPK, die struktuur, funksionaliteit en soekenjins daarvan, ensovoorts. Die verskillende studies wat oor die jare uitgevoer is om die waarde, toepaslikheid en effektiwiteit van die EGPK rakende die outomatiese of semi-outomatiese herkenning, etikettering en onttrekking van terminologie, die samestelling van terminologiese inskrywings asook inskrywings vir die *English–Georgian Dictionary* en die *Georgian–English Learner's Dictionary*, ens., te bepaal en te verbeter, word in die artikel uiteengesit.

Daar word spesifiek klem gelê op die werklike of potensiële toepaslikheid van die korpus vir die leksikografiese aktiwiteite en masjienvertalingsprojekte. Die bevindings van die studie mag ook van waarde wees vir ander hulpbronskaars tale soos Georgies.

Sleutelwoorde: PARALLELLE KORPUS, TERMINOLOGIESE INSKRYWINGS, ENGELS–GEORGIESE WOORDEBOEK, GEORGIES–ENGELSE WOORDEBOEK

1. History of English–Georgian Lexicography

The English–Georgian Parallel Corpus was primarily created for the *Comprehensive English–Georgian Dictionary*, in order to enrich it with entries, corpus illustrative phrases and sentences, and terminological entries. Therefore, in this chapter we will present a brief overview of English–Georgian lexicography.

The history of English–Georgian lexicography in Georgia begins in the 20th century, although there was interest of English authors towards the Georgian and its sister languages in the 18th and the 19th centuries (Margalitadze and Tchighladze 2022; Kikvidze and Pachulia 2019; Margalitadze and Odzeli 2019).

The first English–Georgian dictionary was published in Georgia in the 1940s. The 20th century saw the publication of two comprehensive dictionaries: the *Comprehensive English–Georgian Dictionary* (editor in chief Tinatin Margalitadze) and the *Comprehensive Georgian–English Dictionary* (editor in chief Donald Rayfield).

The work on the *Comprehensive English–Georgian Dictionary* (CEGD) started in the 1970s at the department of English Philology of Ivanè Javakhishvili Tbilisi State University. In the 1980s, a small team of editors embarked upon the mission of fundamentally revising, expanding and updating the dictionary in order to prepare it for publication. In the 1990s the editorial team of the dictionary started digitalization of the dictionary material and in 1995 the printed publi-

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1701 (Article)

cation of the *Comprehensive English–Georgian Dictionary* began in fascicles, on letter-by-letter basis. In 2010, the online version of the dictionary (110 000 entries) was uploaded to the Internet (CEGOD). The primary purpose of the creation of the dictionary was to facilitate the translation of English literature (both belles-lettres or fiction and specialist literature) into Georgian. This is why the dictionary includes contemporary English vocabulary, as well as obsolete, archaic words and meanings and specialist terms (Margalitadze 2012).

The *Comprehensive Georgian–English Dictionary* under editorship of Donald Rayfield was published in London in 2006 by Garnett Press (CGED). Donald Rayfield is an outstanding British Slavist and Kartvelologist. He is the author of a number of monographs on the Russian and Georgian literature. He is also a skilful translator, translating pieces of Georgian literature into English. The *Comprehensive Georgian–English Dictionary* includes contemporary, as well as Old Georgian vocabulary, the word-stock of the Georgian dialects and related Kartvelian languages, and terms from specific branches of knowledge. Donald Rayfield's dictionary contains 140 000 Georgian words and is published in two volumes.

2. English–Georgian Parallel Corpora

There are several English–Georgian parallel corpora, which were mainly developed in the context of multilingual data mining through the Web and have been processed in different ways. Three corpora are presented in this chapter as examples: CCAligned v1, CCAligned v1 and TED2020 v1. The first two are among the largest corpora in number of Georgian data, while the third parallel corpus contains translations of spoken Georgian.

CCAligned v1,[1] "A Massive Collection of Cross-lingual Web-Document Pairs" consists of parallel or comparable web-document pairs in 137 languages aligned with English. The analysis of the automatically translated English–Georgian sentence pairs reveals massive problems of alignment and translation in the Georgian part of the corpus.

Wikimedia v20210402. Wikipedia translations are published by the Wikimedia foundation and their translation system[2] (Tiedemann 2012). The WiKi-Parallel corpus contains 306 languages, including Georgian. The total number of tokens is 918.05M and total number of sentence fragments — 31.62M.

TED2020 v1.[3] This parallel corpus is interesting as it represents a spoken language and was translated by volunteers. This dataset contains a crawl of nearly 4000 TED and TED-X transcripts from July 2020 (Reimers and Gurevych 2020). The transcripts have been translated to more than 100 languages by a global community of volunteers. The parallel corpus contains 108 languages, including Georgian. The total number of tokens — 173.40M, total number of sentence fragments — 11.46M.

The study of above-mentioned, as well as other parallel corpora with the Georgian language reveals that the web-based and automatically created par-

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1701 (Article)

allel corpora have a high rate of linguistic and formatting errors of all types, particularly in a language like Georgian, which is characterized by a complex morphology (Gippert 2016; Harris 1991). For example, the whole parallel corpus of 62 languages — OpenSubtitles (Lison and Tiedemann 2016) is completely unusable for Georgian due to the formatting and coding errors.

2.1 English–Georgian Parallel Corpus of Ilia State University

The work on the EGPC started in 2011. The corpus consists of two sub-corpora: the sub-corpus of scientific and domain-specific texts and the sub-corpus of fiction (translated from Georgian into English and vice versa). From the very beginning of the project the decision was made to concentrate on the quality of translated texts, as well as the structuring of the data in it, as the primary goal of developing the EGPC was its application in English–Georgian lexicography.

The most important part of the sub-corpus of scientific texts constitute translations of professor Arrian Tchanturia, a prominent Georgian scholar, editor, translator and lexicographer (member of editorial boards of both comprehensive dictionaries: English–Georgian and Georgian–English). He was one of the first scholars to start translation of Georgian scholarly and scientific literature into English from the 1960s. His translation legacy includes hundreds of pages of translated abstracts, papers, and books from Georgian into English covering practically all fields of knowledge. The desire to transform this legacy into an English–Georgian Parallel Corpus and to apply it in the work on the CEGD gave the impetus to the development of this project (Margalitadze 2014). Later this sub-corpus was extended with other translations and grew into a sub-corpus of scientific and domain-specific texts. At the next stage, translations of literary works were added to the corpus.

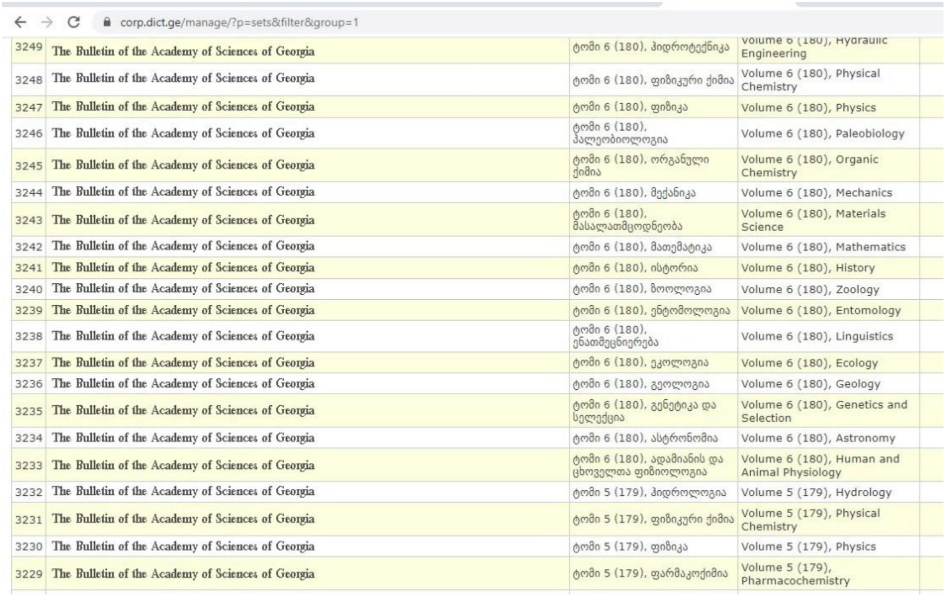
2.2 The Structure of the English–Georgian Parallel Corpus

The principles of arrangement of data in the corpus databases were worked out after a long period of deliberation and aimed at the arrangement of texts in databases in a way that would enable the application of the corpus in general and specialized lexicography in future. The platform is based on the program created for the English–Hungarian parallel corpus 'HunAlign freeware tool'.[4]

The structure of the database consists of three sections: text groups, text sets and sentence pairs. Each text group is subdivided into text sets and each text set is further subdivided into sentence pairs. Text group is the largest unit of the database and it consists of a variety of texts. At the present moment the EGPC comprises over 70 text groups of different sizes and new material is added to the corpus on a daily basis.

One of the largest text groups in the sub-corpus of scientific texts is *The Bulletin of the Academy of Sciences of Georgia*. It incorporates material from issues published over a period of 24 years. This material consists of English–Georgian abstracts of scientific papers from virtually all fields of knowledge. This subcorpus also includes scholarly bilingual papers published in several bilingual scholarly journals in Georgia, e.g. Kartvelology and Kadmos. One of the text groups represents a series of publications about important archaeological excavations in Georgia. Text groups also include scholarly books, manuals of different subjects translated from English into Georgian, materials published by the Legislative Herald of Georgia, election administration, the Government of Georgia, and materials collected from different websites.

Each text group, as mentioned above, is subdivided into text sets. Text sets vary according to the type of the text group. E.g., the text group *The Bulletin of the Academy of Sciences of Georgia* is divided into volumes (with each volume containing three issues) and each volume (text set) contains abstracts of one domain: volume 6 (180) ecology; volume 6 (180) entomology; volume 6 (180) geology; volume 6 (180) human and animal physiology; volume 6 (180) mechanics; volume 6 (180) organic chemistry, etc. (see Figure 1).



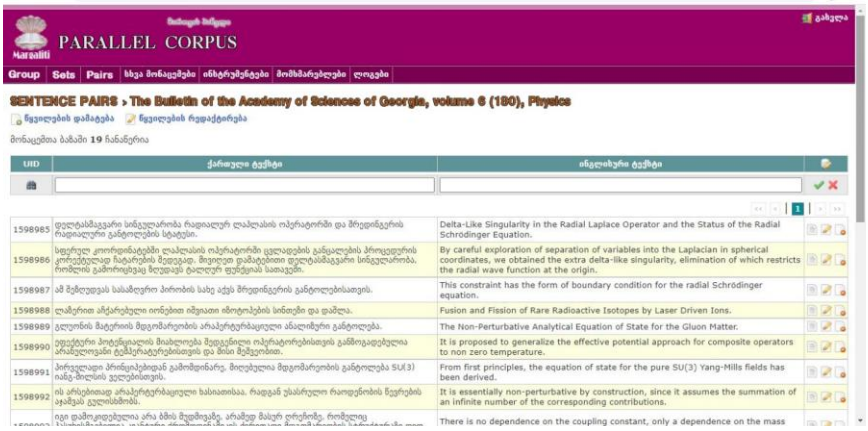
corp.dict.ge/manage/?p=sets&filter&group=1

3249	The Bulletin of the Academy of Sciences of Georgia	ტომი 6 (180), ჰიდროტექნიკა	Volume 6 (180), Hydraulic Engineering	
3248	The Bulletin of the Academy of Sciences of Georgia	ტომი 6 (180), ფიზიკური ქიმია	Volume 6 (180), Physical Chemistry	
3247	The Bulletin of the Academy of Sciences of Georgia	ტომი 6 (180), ფიზიკა	Volume 6 (180), Physics	
3246	The Bulletin of the Academy of Sciences of Georgia	ტომი 6 (180), პალეობიოლოგია	Volume 6 (180), Paleobiology	
3245	The Bulletin of the Academy of Sciences of Georgia	ტომი 6 (180), ორგანული ქიმია	Volume 6 (180), Organic Chemistry	
3244	The Bulletin of the Academy of Sciences of Georgia	ტომი 6 (180), მექანიკა	Volume 6 (180), Mechanics	
3243	The Bulletin of the Academy of Sciences of Georgia	ტომი 6 (180), მასალათმცოდნეობა	Volume 6 (180), Materials Science	
3242	The Bulletin of the Academy of Sciences of Georgia	ტომი 6 (180), მათემატიკა	Volume 6 (180), Mathematics	
3241	The Bulletin of the Academy of Sciences of Georgia	ტომი 6 (180), ისტორია	Volume 6 (180), History	
3240	The Bulletin of the Academy of Sciences of Georgia	ტომი 6 (180), ზოოლოგია	Volume 6 (180), Zoology	
3239	The Bulletin of the Academy of Sciences of Georgia	ტომი 6 (180), ენტომოლოგია	Volume 6 (180), Entomology	
3238	The Bulletin of the Academy of Sciences of Georgia	ტომი 6 (180), ენათმეცნიერება	Volume 6 (180), Linguistics	
3237	The Bulletin of the Academy of Sciences of Georgia	ტომი 6 (180), ეკოლოგია	Volume 6 (180), Ecology	
3236	The Bulletin of the Academy of Sciences of Georgia	ტომი 6 (180), გეოლოგია	Volume 6 (180), Geology	
3235	The Bulletin of the Academy of Sciences of Georgia	ტომი 6 (180), გენეტიკა და სელექცია	Volume 6 (180), Genetics and Selection	
3234	The Bulletin of the Academy of Sciences of Georgia	ტომი 6 (180), ასტრონომია	Volume 6 (180), Astronomy	
3233	The Bulletin of the Academy of Sciences of Georgia	ტომი 6 (180), ადამიანის და ცხოველთა ფიზიოლოგია	Volume 6 (180), Human and Animal Physiology	
3232	The Bulletin of the Academy of Sciences of Georgia	ტომი 5 (179), ჰიდროლოგია	Volume 5 (179), Hydrology	
3231	The Bulletin of the Academy of Sciences of Georgia	ტომი 5 (179), ფიზიკური ქიმია	Volume 5 (179), Physical Chemistry	
3230	The Bulletin of the Academy of Sciences of Georgia	ტომი 5 (179), ფიზიკა	Volume 5 (179), Physics	
3229	The Bulletin of the Academy of Sciences of Georgia	ტომი 5 (179), ფარმაკოქიმია	Volume 5 (179), Pharmacochemistry	

Figure 1

Other text groups are structured differently. Scientific and scholarly journals are divided into text sets according to separate articles; books are divided into chapters and so on. Such organization of the database allows the sorting of the material according to domains as well as many other criteria.

Text sets are further subdivided into sentence pairs. These are aligned English–Georgian parallel sentences (see Figure 2).



SENTENCE PAIRS › The Bulletin of the Academy of Sciences of Georgia, volume 6 (180), Physics

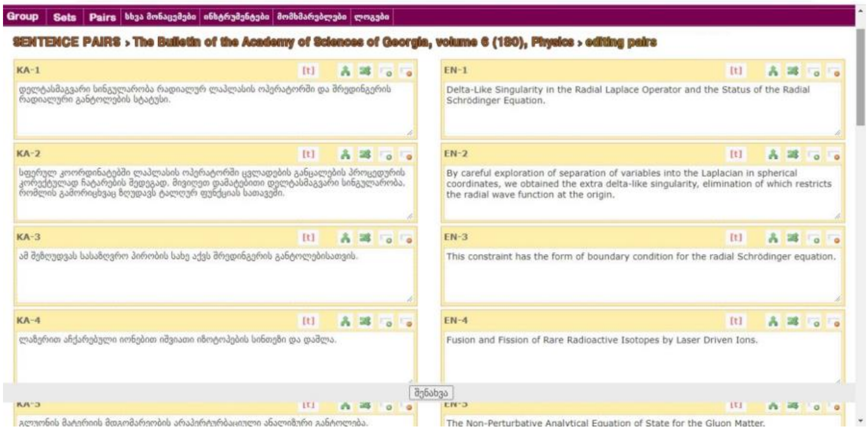
მონაცემთა ბაზაში 19 ჩანაწერია

UID	ქართული ტექსტი	ინგლისური ტექსტი
1598985	დელტასმაგვარი სინგულარობა რადიალურ ლაპლასის ოპერატორში და შრედინგერის რადიალური განტოლების სტატუსი.	Delta-Like Singularity in the Radial Laplace Operator and the Status of the Radial Schrödinger Equation.
1598986	სფერულ კოორდინატებში ლაპლასის ოპერატორში ცვლადების განცალების პროცედურის კორექტულად ჩატარების შედეგად. მივიღეთ დამატებითი დელტასმაგვარი სინგულარობა, რომლის გამორიცხვაც ზღუდავს ტალღურ ფუნქციას სათავეში.	By careful exploration of separation of variables into the Laplacian in spherical coordinates, we obtained the extra delta-like singularity, elimination of which restricts the radial wave function at the origin.
1598987	ამ შეზღუდვას სასაზღვრო პირობის სახე აქვს შრედინგერის განტოლებისათვის.	This constraint has the form of boundary condition for the radial Schrödinger equation.
1598988	ლაზერით აჩქარებული იონებით იშვიათი იზოტოპების სინთეზი და დაშლა.	Fusion and Fission of Rare Radioactive Isotopes by Laser Driven Ions.
1598989	გლუონის მატერიის მდგომარეობის არაპერტურბაციული ანალიზური განტოლება.	The Non-Perturbative Analytical Equation of State for the Gluon Matter.
1598990	ეფექტური პოტენციალის მიახლოება შედგენილი ოპერატორებისთვის განზოგადებულია არანულოვანი ტემპერატურებისთვის და მისი შეშვებით.	It is proposed to generalize the effective potential approach for composite operators to non zero temperature.
1598991	პირველადი პრინციპებიდან გამომდინარე, მიღებულია მდგომარეობის განტოლება SU(3) იანგ-მილსის ველებისთვის.	From first principles, the equation of state for the pure SU(3) Yang-Mills fields has been derived.
1598992	ის არსებითად არაპერტურბაციული ხასიათისაა, რადგან უსასრულო რაოდენობის წევრების აჯამვას გულისხმობს.	It is essentially non-perturbative by construction, since it assumes the summation of an infinite number of the corresponding contributions.
	იგი დამოკიდებულია არა ბმის მუდმივაზე, არამედ მასურ ღრეჩოზე, რომელიც	There is no dependence on the coupling constant, only a dependence on the mass

Figure 2

Text sets are uploaded to the special fields in the database, allocated to English and Georgian.

The program automatically breaks down text sets into sentence pairs (see Figure 3).



SENTENCE PAIRS › The Bulletin of the Academy of Sciences of Georgia, volume 6 (180), Physics › editing pairs

KA-1
დელტასმაგვარი სინგულარობა რადიალურ ლაპლასის ოპერატორში და შრედინგერის რადიალური განტოლების სტატუსი.

EN-1
Delta-Like Singularity in the Radial Laplace Operator and the Status of the Radial Schrödinger Equation.

KA-2
სფერულ კოორდინატებში ლაპლასის ოპერატორში ცვლადების განცალების პროცედურის კორექტულად ჩატარების შედეგად. მივიღეთ დამატებითი დელტასმაგვარი სინგულარობა, რომლის გამორიცხვაც ზღუდავს ტალღურ ფუნქციას სათავეში.

EN-2
By careful exploration of separation of variables into the Laplacian in spherical coordinates, we obtained the extra delta-like singularity, elimination of which restricts the radial wave function at the origin.

KA-3
ამ შეზღუდვას სასაზღვრო პირობის სახე აქვს შრედინგერის განტოლებისათვის.

EN-3
This constraint has the form of boundary condition for the radial Schrödinger equation.

KA-4
ლაზერით აჩქარებული იონებით იშვიათი იზოტოპების სინთეზი და დაშლა.

EN-4
Fusion and Fission of Rare Radioactive Isotopes by Laser Driven Ions.

KA-5
გლუონის მატერიის მდგომარეობის არაპერტურბაციული ანალიზური განტოლება.

EN-5
The Non-Perturbative Analytical Equation of State for the Gluon Matter.

Figure 3

At the next stage, the sentences broken down automatically are manually aligned with the help of tools provided at the top right corner of each block. These tools allow one to add or delete blocks or to exchange places between two blocks. Manual alignment usually corrects minor errors, e.g. cases when one English sentence is translated by two Georgian sentences or vice versa. The result of this approach is high-quality, ideally aligned sentence pairs.

Texts uploaded to the sub-corpus of scientific texts comprise all fields of knowledge: mathematics, mechanics, geophysics, chemistry, hydrology, geol-

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1701 (Article)

ogy, palaeontology, machine building science, hydraulic engineering, electrical engineering, botany, genetics, physiology, biophysics, biochemistry, entomology, experimental morphology, experimental medicine, financing, archaeology, ethnography, Kartvelology etc. The sub-corpus of fiction contains translations of Georgian belles-lettres into English, as well as translations of English authors into Georgian. The sub-corpus of fiction also includes translations of plays.

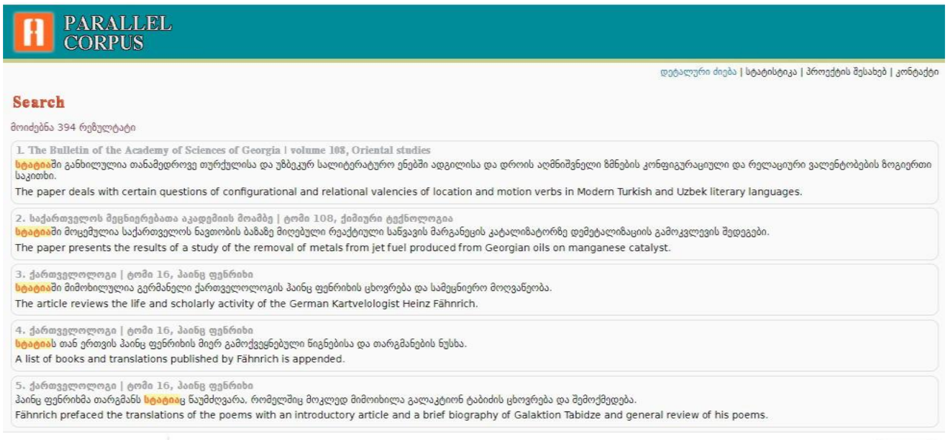
At present, the corpus contains up to 70 text groups, 5 000 text sets, 400 000 manually aligned sentence pairs and 7 million tokens. The EGPC has an interface for searching Georgian or English words and collocations and displaying the proper text pairs containing the search results on the screen. Each sentence pair is numbered and is supplied with the information about corresponding text group and text set (see Figure 4).

Thus, unlike the English–Georgian parallel corpora, discussed in chapter 2, the EGPC of Ilia State University is characterized by the following features:

(1) high-quality translations edited by human specialists,
(2) accurate and error-free alignment of sentences, and
(3) constantly growing corpus through parallel use of human specialists and NLP.

On all three points, the *Comprehensive English–Georgian Dictionary* acts as a lexicographic source of the translation quality.

When the corpus reached 4 million tokens, studies were conducted for evaluating the efficiency of the Corpus for English–Georgian Lexicography. Three main tasks were identified for the EGPC: compiling terminological entries, compiling entries for the English–Georgian Dictionary and compiling entries for the Georgian–English Learner's Dictionary. These studies were carried out within the framework of MA and PhD programmes in lexicography with the active participation of MA and PhD students in lexicography.



მოიძებნა 394 რეზულტატი

1. The Bulletin of the Academy of Sciences of Georgia | volume 108, Oriental studies
სტატიაში განხილულია თანამედროვე თურქულისა და უზბეკურ სალიტერატურო ენებში ადგილისა და დროის აღმნიშვნელი ზმნების კონფიგურაციული და რელაციური ვალენტობების ზოგიერთი საკითხი.
The paper deals with certain questions of configurational and relational valencies of location and motion verbs in Modern Turkish and Uzbek literary languages.

2. საქართველოს მეცნიერებათა აკადემიის მოამბე | ტომი 108, ქიმიური ტექნოლოგია
სტატიაში მოცემულია საქართველოს ნავთობის ბაზაზე მიღებული რეაქტიული საწვავის მარგანეცის კატალიზატორზე დემეტალიზაციის გამოკვლევის შედეგები.
The paper presents the results of a study of the removal of metals from jet fuel produced from Georgian oils on manganese catalyst.

3. ქართველოლოგი | ტომი 16, ჰაინც ფენრიხი
სტატიაში მიმოხილულია გერმანელი ქართველოლოგის ჰაინც ფენრიხის ცხოვრება და სამეცნიერო მოღვაწეობა.
The article reviews the life and scholarly activity of the German Kartvelologist Heinz Fähnrich.

4. ქართველოლოგი | ტომი 16, ჰაინც ფენრიხი
სტატიას თან ერთვის ჰაინც ფენრიხის მიერ გამოქვეყნებული წიგნებისა და თარგმანების ნუსხა.
A list of books and translations published by Fähnrich is appended.

5. ქართველოლოგი | ტომი 16, ჰაინც ფენრიხი
ჰაინც ფენრიხმა თარგმანს სტატიაც წაუმძღვარა, რომელშიც მოკლედ მიმოიხილა გალაკტიონ ტაბიძის ცხოვრება და შემოქმედება.
Fähnrich prefaced the translations of the poems with an introductory article and a brief biography of Galaktion Tabidze and general review of his poems.

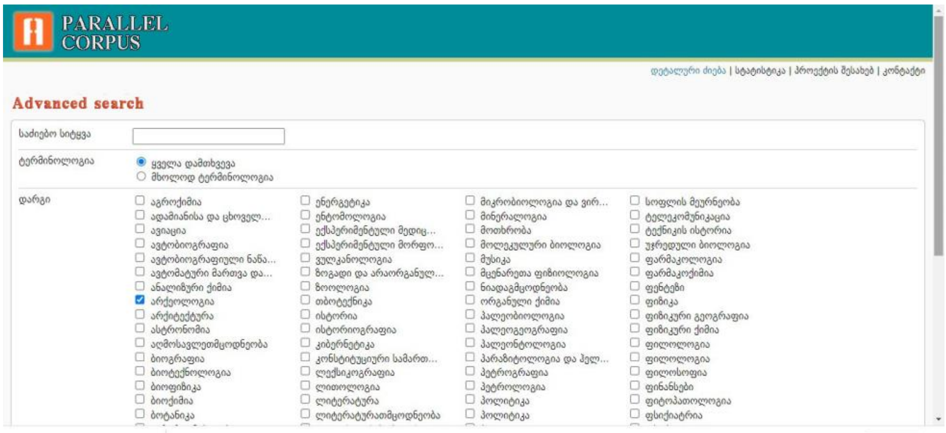
Figure 4

2.3 Application of the English–Georgian Parallel Corpus in Terminology

The work on the elaboration of the methodology of tagging and extracting specialized terminology from the corpus started in 2015. A special module, the terminological module, was developed that allows the extraction of the previously tagged terminology from the corpus. After the development of this module, the function "Recognition of and search for the tagged terms in the corpus" was added to the existing functions of the corpus control panel, namely:

— Management functionalities of text groups
— Management functionalities of text sets
— Management functionalities of text pairs
— Automatic breakdown of texts by sentences, sentence alignment, generation of pairs and further manual alignment options.

An advanced search function was added to the simple search functionality of the EGPC. Figure 5 shows the advanced search page which displays all fields of knowledge represented by texts of different sizes in the EGPC: aviation, archaeology, architecture, oriental studies, botany, zoology, biology, geology, ecology, ethnography, economics, banking, history, Kartvelian studies, hydrology, psychology and many others. The principles of the arrangement of corpus databases into text groups and text sets, described above, allow one to sort terminology according to domains and to extract them from the corpus for further lexicographic processing. Specialized terms are extracted from the corpus alongside their English equivalents and, significantly, collocations of terms with their respective English translations can also be extracted.



PARALLEL CORPUS

დეტალური ძიება | სტატისტიკა | პროექტის შესახებ | კონტაქტი

Advanced search

საძიებო სიტყვა

ტერმინოლოგია
● ყველა დამთხვევა
○ მხოლოდ ტერმინოლოგია

დარგი
□ აგროქიმია
□ ადამიანისა და ცხოველ...
□ ავიაცია
□ ავტობიოგრაფია
□ ავტობიოგრაფიული ნაწა...
□ ავტომატური მართვა და...
□ ანალიზური ქიმია
☑ არქეოლოგია
□ არქიტექტურა
□ ასტრონომია
□ აღმოსავლეთმცოდნეობა
□ ბიოგრაფია
□ ბიოტექნოლოგია
□ ბიოფიზიკა
□ ბიოქიმია
□ ბოტანიკა
□ ენერგეტიკა
□ ენტომოლოგია
□ ექსპერიმენტული მედიც...
□ ექსპერიმენტული მორფო...
□ ვულკანოლოგია
□ ზოგადი და არაორგანულ...
□ ზოოლოგია
□ თბოტექნიკა
□ ისტორია
□ ისტორიოგრაფია
□ კიბერნეტიკა
□ კონსტიტუციური სამართ...
□ ლექსიკოგრაფია
□ ლითოლოგია
□ ლიტერატურა
□ ლიტერატურათმცოდნეობა
□ მიკრობიოლოგია და ვირ...
□ მინერალოგია
□ მოთხრობა
□ მოლეკულური ბიოლოგია
□ მუსიკა
□ მცენარეთა ფიზიოლოგია
□ ნიადაგმცოდნეობა
□ ორგანული ქიმია
□ პალეობიოლოგია
□ პალეოგეოგრაფია
□ პალეონტოლოგია
□ პარაზიტოლოგია და ჰელ...
□ პეტროგრაფია
□ პეტროლოგია
□ პოლიტიკა
□ პოლიტიკა
□ სოფლის მეურნეობა
□ ტელეკომუნიკაცია
□ ტექნიკის ისტორია
□ უჯრედული ბიოლოგია
□ ფარმაკოლოგია
□ ფარმაკოქიმია
□ ფენტეზი
□ ფიზიკა
□ ფიზიკური გეოგრაფია
□ ფიზიკური ქიმია
□ ფილოლოგია
□ ფილოლოგია
□ ფილოსოფია
□ ფინანსები
□ ფიტოპათოლოგია
□ ფსიქიატრია

Figure 5

The analysis of terminological entries created on the basis of the EGPC revealed that the corpus is a very efficient source for the CEGOD and that it can enrich the dictionary with terminology of different domains. Two cases are to be noted: some terms were not recorded in the CEGOD and were added to it from the corpus, and in some cases terminological entries of the CEGOD were improved by adding new collocations to them. For example, the financial term *direct debit* was introduced in the CEGOD with the following collocations and their Georgian translations: *direct debit order, direct debit service, direct debit transfer*. The financial terms *documentary collection* and *encashment order* were added to the dictionary macrostructure. The economic term *inflation* had been already included in the CEGOD, but the corpus material enabled the addition of the following collocations: *high inflation, the rate of inflation, high rate of inflation, a period of inflation, demand-pull inflation, cost-push inflation, to reduce the threat of inflation.* These collocations are supplied with Georgian translations from the corpus. The following collocations and their Georgian equivalents were added to the economic term *cost*: *production costs, operating costs, fixed costs, variable costs, to increase/raise costs, to reduce costs, to cut costs, rising costs, marginal costs, external costs, shipping costs, refining costs, to incur costs*.

The EGPC can also be applied in English–Georgian terminological dictionary projects, but only as one of the sources. It is unlikely to have enough translations of specialized texts in one domain to fully rely only on the parallel corpus while compiling a bilingual dictionary of one field of knowledge.

One of the recent studies conducted in the EGPC was the testing of different tools for automatic or semi-automatic recognition, tagging and extraction of terminology from the parallel corpus. Different tools were tested for this purpose, but the most efficient one proved to be *Synchroterm*, developed by a Canadian computer program company Terminotix.[5] The study will continue in this direction and the selected program will be integrated with the EGPC in order to facilitate work on the terminology.

2.4 Application of the English–Georgian Parallel Corpus for Georgian–English Learner's Dictionary

Compilation of Georgian–English Learner's Dictionary (GELD) is high on the agenda of the Centre for Lexicography and Language Technologies. The *Comprehensive Georgian–English Dictionary*, published under the general editorship of D. Rayfield, is mostly aimed at foreign scholars interested in Georgian and its sister languages, mediaeval Georgian literature, and the history of Georgia in the Middle Ages, when this country played an important role in European history. Proceeding from these considerations, the macrostructure of the dictionary includes Old and Middle Georgian words and dialectal material, which is important for the main target group of the CGED. The dictionary is more concerned with the macrostructure, reflected in the number of entries (140 000).

On the other hand, Georgian learners of English need more information about the usage of Georgian words and their rendition in English. In other words, they need a dictionary which is oriented on text synthesis, text production, speaking/writing and not only text analysis, i.e. understanding spoken/ written text. Our decades-long experience of working on the CEGD has revealed that there is considerable semantic asymmetry between the English and Georgian languages. As a result, an English word cannot always be translated by one Georgian equivalent in various contexts and often needs different contextual equivalents to properly translate its meaning. In the CEGD our editorial team introduced two levels of equivalence in an entry: meaning equivalence and contextual/translation equivalence, which is discussed in detail in our paper presented at the XVII International Congress of EURALEX (Margalitadze and Meladze 2016). Therefore, illustrative phrases and sentences, which show the usage of an English word and its Georgian translations, are important in the CEGD entries. This is also true for the reverse Georgian–English dictionary: Georgian words should be supplied with different illustrative phrases, sentences and collocations translated into English. These considerations determined our interest in the EGPC and its efficiency for the GELD project.

The study of the effectiveness of the EGPC for the compilation of the GELD entries yielded very positive results. In many cases, the data collected from the corpus enabled editors to produce adequate dictionary entries and to identify and single out polysemous meanings of Georgian words, sometimes even more meanings than are registered in monolingual dictionaries of Georgian. The corpus data provides many illustrative phrases, collocations and sentences for Georgian words with their respective English equivalents.

For example, for the Georgian word *მტკიცე mṭḳice* two polysemous meanings are identified and each meaning is well-illustrated with the corpus examples:

> *მტკიცე mṭḳice* **1.** (*firm, solid, steady*) მტკიცე ავეჯი solid furniture; მტკიცე ქიმიური ბმები firm chemical bonds; ფანჯარა ძალიან მტკიცე მინისგან არის დამზადებული the window is made from very strong glass; განა შეიძლება შედეგი მტკიცე იყოს? Can the result be sound?; მტკიცე ფეხსაცმელი durable shoes; მტკიცე ნივთიერება enduring substance; მტკიცე და დაუძლეველი ზღუდე fast and impassable barrier; მტკიცე კარები a solid door; **2.** (*determined, decisive, resolute*) მტკიცე ხასიათი decisive character; მტკიცე ტრადიცია deep-seated tradition; მტკიცე ფასი determined price; მტკიცე ნების ადამიანი a man of hard, unbending will; მტკიცე ნებისყოფის ქონა to have an iron will; მტკიცე ოპტიმისტი a resolute optimist; მტკიცე ბიუროკრატიული კონტროლი rigid bureaucratic controls; მონარქიის მტკიცე მხარდამჭერი a staunch supporter of the monarchy; მტკიცე ნაბიჯებით with sure steps.

The corresponding entry from D. Rayfield's CEGD presents the same Georgian word in the following way:

1. Solid, firm; established; მტკიცე ნაბიჯი a decisive step; მტკიცე უარი a definite no;

2. Of good cheer (*this is an obsolete meaning of this adjective which will not be presented in a learner's dictionary*).

The English language abounds in synonyms. For a Georgian learner of English, it is important to know which synonym should be used in a particular context. From this point of view the EGPC provides really useful and important data about usage of Georgian words and, even more important, their translations into English.

For the Georgian verb *დაფარვა daparva* the corpus data singles out four meanings:

> *დაფარვა daparva* **1.** (*to cover*) მტვრით ხარ დაფარული you are covered in dust; მიწა თოვლით იყო დაფარული the ground was blanketed with snow; **2.** (*to keep secret, to conceal*) შიშის [მღელვარების, ნერვიულობის] დაფარვა to conceal one's fear [excitement, nervousness]; სიმართლის დაფარვა to hide the truth; მტრული დამოკიდებულების დაფარვა მეგობრობის ნიღბით to mask one's enmity under an appearance of friendliness; **3.** (*to pay debt, to compensate*) სესხის დაფარვა to pay a loan; მან ვალი დაფარა he wiped out the debt; **4.** (*to protect, to defend*) სამშობლოს მტრისგან დაფარვა to defend one's homeland from an enemy; თვალების მზისგან დაფარვა to protect one's eyes from the sun; ◊ **დაფარვის ზონა** coverage area.

The corresponding entry from D. Rayfield's CGED presents the same four meanings without providing examples of usage:

1. Covering (*with snow, clothes*); დაფარვის ზონა (*mobile phone, etc.*) coverage area;
2. Keeping hidden;
3. Paying off (debt);
4. Defence.

At present, the work is underway on the issues connected with the automation of data collection from the corpus in order to facilitate the work of lexicographers.

2.5 Application of the English–Georgian Parallel Corpus for the Comprehensive English–Georgian Dictionary

Further studies included the assessment of the corpus's efficacy for the *Comprehensive English–Georgian Dictionary*. Our aim was to assess the volume and representativeness of the EGPC by means of looking up and retrieving corpus data with respect to some pre-selected lexical units. This would enable us to find out to what extent the polysemy of these words was traceable in the parallel English–Georgian sentences represented in the corpus, and how helpful the data retrievable from the corpus could be for the composition of more or less full-fledged dictionary articles.

To that end, we chose a number of nouns, verbs, adjectives and adverbs.

Context-based meanings retrieved from the database permitted the composition of dictionary entries with some considerable scope of polysemy.

Before proceeding to general conclusions, we would like to demonstrate the material with respect to the lexical unit *dream* (noun + verb) that was extracted from the corpus. This article is a characteristic example of dictionary articles based on the data retrieved from the EGPC:

> dream *noun* **1.** (*a vision during sleep*) სიზმარი; for a long time, I could not shut my eyes and, when I did get to sleep, I was transported by dreams დიდხანს თვალი ვერ დავხუჭე, და, რომ დამეძინა, სიზმრებმა წამიღეს; **2.** (*an aspiration, a wish to have or be something*) ოცნება; his entire poetry clearly expresses the dreams and aspirations of the Georgian people მთელი მისი პოეზია ქართველი ხალხის ოცნებებისა და მისწრაფებების ნათლად გამომხატველია; **3.** (*daydream, reverie*) ზმანება; the tender, sweet dream of a love seen once ოდესღაც ნანახი სატრფოს ნაზი, ტკბილი ზმანება; now he could know that this had truly happened and was not a dream ახლა საბოლოოდ დარწმუნდა, რომ ეს ყველაფერი ზმანება კი არა, ცხადი იყო; life is a dream სიცოცხლე ზმანებაა.
>
> dream *verb* **1.** (*to experience a dream during sleep*) დასიზმრება (<და>ესიზმრება); "is this the man I dreamt of?" she worried "ნუთუ ის კაცია, ვინც დამესიზმრაო" - წუხდა ქალი; **2.** (*to have a deep aspiration or hope*) ოცნება (ოცნებობს); the point is that many crusaders dreamed of seizing lands and becoming rich საქმე ისაა, რომ ბევრი ჯვაროსანი მიწების ხელში ჩაგდებასა და გამდიდრებაზე ოცნებობდა; he dreams of creating a library and setting up a printing press იგი ოცნებობს ბიბლიოთეკის შექმნასა და სტამბის დაარსებაზე; **3.** (*to daydream, to pass time in reverie*) ხილვის / ზმანების ქონა (აქვს); *რაიმე* ეზმანება; he only dreamed of foreign lands now and of the lions on the beach მას ახლა მხოლოდ უცხო მხარე და სანაპიროზე გამოფენილი ლომები ეზმანებოდა; **4.** (*to regard something as feasible or practical, to imagine*) უარყოფით წინადადებებში: ფიქრი (ფიქრობს), განზრახვა; the French will never dream of it ფრანგებს ეს არც დაესიზმრებათ; "I could never dream of such success in my own country," she admitted frankly "ჩემს სამშობლოში ამგვარი წარმატება არც კი დამესიზმრებოდაო" - აღიარა მან გულწრფელად.

The above entries (DREAM *noun* + *verb*) provide some interesting information about the subject under discussion. Comparing these entries with those included in the *Comprehensive English–Georgian Dictionary* (https://dictionary.ge/ka/word/dream+I/ and https://dictionary.ge/ka/word/ dream+II/) we could see that many polysemous meanings present in the entries of CEGD can be seen in corpus-based entries as well. Moreover, the third verbal meaning '*to daydream, to pass time in reverie*', is absent in the CEGD, while the same meaning could be identified based on the contexts attested in the parallel sentences retrieved from the corpus.

On the other hand, some meanings, e.g. '*to dream up*' (to invent, concoct) which is included in the entry of the *Comprehensive English–Georgian Dictionary*, is absent from our corpus-based entry, as far as no sentences/contexts, where 'to dream (up)' would denote 'inventing or concocting something', could have been retrieved from the EGPC.

Meanwhile, the further analysis of the dictionary entries, composed using the data retrieved from the corpus, showed that some meanings of polysemous words had more hits in the corpus, while other ones were very scarce and only few occurrences thereof could be attested in the corpus database. For instance, in the case of the adjective *short*, we obtained many contexts, where *short* meant 'not lengthy', 'of short duration' or 'deficient in something' or 'lacking something', but (somewhat surprisingly), there were very few cases were *short* meant 'not long', and only one case where *short* referred to the human stature (i.e., meaning 'not high or tall'). Only one result for *short* with its semantic value referring to vowel shortness *v* length (in prosody and phonetics) came as no surprise, while the scarceness of the contexts with *short* meaning 'not long' or 'not high/tall' required some explanation. Our best guess is that a relatively large proportion of purely scientific or official texts in our corpus (*The Bulletin of the Academy of Sciences of Georgia*, legislative documents, texts related to the economic, financial and banking activities, etc.) may somehow account for the relatively scarce representation of words (*short* in this particular case) with semantic values related to everyday life and 'ordinary' situational contexts.

To summarize, we can state that our investigation has allowed us to arrive at certain conclusions. Since Georgian, as a language, is under-resourced and lacks large amounts of parallel Georgian–English texts, we cannot expect the EGPC to yield data for comprehensive dictionaries with full-size entries based on extensive polysemy. Furthermore, since approximately two thirds of the texts included in our corpus are those translated from Georgian into English, the application of the corpus-based data extracted from the corpus seems to be more appropriate for *Georgian–English Learner's Dictionary* project. It should be also mentioned that even at the present stage, the corpus proves to be very useful source for enriching the CEGD entries with additional senses or good dictionary examples. This study also showed that the development of the corpus should concentrate on texts translated from English into Georgian to provide balance and have an equal proportion of texts translated from Georgian into English and vice versa. The corpus also needs to be balanced by including more translations of literary works as opposed to translations of scientific and official texts.

3. Application of the English–Georgian Parallel Corpus for English–Georgian/Georgian–English Machine Translation Project

In 2018 our editorial team realized that we possessed the data that could be instrumental in Georgian–English/English–Georgian machine translation project (Margalitadze and Pourtskhvanidze 2019). Such a project needs: (a) a collection of software platforms and models adapted to the specifics of the Georgian language, and (b) professionally translated English–Georgian parallel sentences in the quantities and amounts as necessary to ensure quality saturation.

As a software prototype for the project, researches based on the simulation of human abilities within the framework of Artificial Intelligence were selected. DeepLearning technology has demonstrated many successful examples of becoming the leading technology and methodological framework. Out of effective models implemented within this framework, machine translation is one of the three most successful examples.

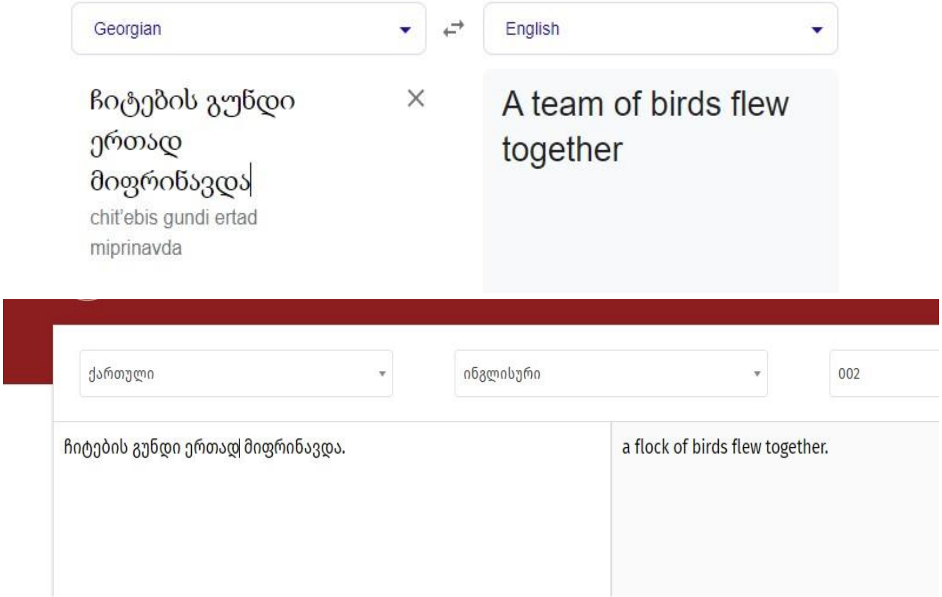
Concerning English–Georgian parallel sentences, our team possesses a database unique for the Georgian language. The base includes two sub-components: the database of the *Comprehensive English–Georgian Dictionary* mentioned above (chapter 1), and the base of the English–Georgian Parallel Corpus, discussed in Chapters 2.1 and 2.2.

For the machine translation project some additional studies were conducted on the corpus in order to evaluate it from the point of view of lexical richness (Kubát and Milička 2013; Brezina 2018). Due to its limitations in terms of digital resources, Georgian needs qualitative processing of data alongside proper structuring of databases. Balancing text types or genres is one such effort. Linguistic diversity in the corpus is represented on the basis of the lexical diversity of its components. The value of lexical diversity was obtained by automatically calculating type-token ratios (TTR) in a text. A clustered calculation for the whole corpus provided the overall picture of equal or unequal distribution of TTR values in the corpus, showing gaps in terms of the balance. Further development of the corpus will take the TTR values into account in the selection of text collections (Margalitadze and Pourtskhvanidze 2021).

At the present moment, the initial stage of the data training for machine translation is over and we are in the process of analysing the first results of the English–Georgian/Georgian–English machine translation program.[6] The training was conducted with 367 000 English–Georgian sentence pairs in which 267 000 pairs were from the EGPC and 100 000 from the CEGD. The data was trained in the OpenNMT model.[7] Although our aim is to reach up to 1 million sentence pairs, the results of this initial stage are very promising. The program has learnt even very specific vocabulary quite well, and deals particularly well with collocations.[8] From this point of view, our machine translation program, in some cases, provides more accurate translations from Georgian into English, than Google translate, which is based on the 1.3 million English–Georgian sentence pairs.[9] Below are quoted some examples which illustrate the difference in the English translations of Georgian sentences by the Google translate and our translator:

1. ღორების კოლტი ზღვაში გადავარდა:
   ġorebis ḳolṭi zġvaši gadavarda
   The Google translate: The pig colt fell into the sea.
   Our translator: A herd of swine fell into the sea.

2. მგლების ხროვა მას ყოველი მხრიდან უტევდა:
   mglebis xrova mas q̇oveli mxridan uṭevda
   The Google translate: A herd of wolves attacked him from all sides.
   Our translator: A pack of wolves was attacking him from all sides.
3. არწივი ცაში ლივლივებდა:
   arc̣ivi caši livlivebda
   The Google translate: The eagle was flying in the sky.
   Our translator: The eagle was soaring in the sky.
4. მდინარე ტყეში მორაკრაკებდა:
   mdinare ṭq̇eši morak̇rak̇ebda
   The Google translate: The river was flowing in the forest.
   Our translator: The river bubbled in the forest.
5. ფარდები ქარში ფრიალებდა:
   pardebi karši prialebda
   The Google translate: The curtains were flying in the wind.
   Our translator: Curtains fluttered in the wind.
6. ჩიტების გუნდი ერთად მიფრინავდა:
   čiṭebis gundi ertad miprinavda
   The Google translate: A team of birds flew together.
   Our translator: A flock of birds flew together.
   (see Figure 6).



Georgian ⇄ English

ჩიტების გუნდი ერთად მიფრინავდა
chit'ebis gundi ertad miprinavda

A team of birds flew together

ქართული ინგლისური 002

ჩიტების გუნდი ერთად მიფრინავდა.

a flock of birds flew together.

Figure 6

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1701 (Article)

4. Conclusion

As described in above chapters, various studies were conducted in order to evaluate the applicability and efficiency of the English–Georgian Parallel Corpus (EGPC) for lexicographical and machine translation projects. These are: (a) the analysis of terminological entries created on the basis of the EGPC, which revealed that the corpus can be a very efficient source for the *Comprehensive English–Georgian Online Dictionary* (CEGOD), enriching the dictionary with terms from different domains; (b) the studies conducted in the EGPC with different tools for automatic or semi-automatic recognition, tagging and extraction of terminology from the corpus; (c) the studies intended to identify the value of the EGPC for compiling entries for *English–Georgian Dictionary* and entries for *Georgian–English Learner's Dictionary*; and (d) the studies for testing the efficacy of the EGPC for machine translation.

The wide range of research activities described above highlight the importance of well-balanced parallel corpora based on adequate, high-quality translations and thoughtfully and meticulously structured data for modern bilingual lexicography. These studies encouraged us to continue the work on the EGPC. The project will develop both quantitatively and qualitatively. From the quantitative point of view the aim is to reach up to 1 million English–Georgian sentence pairs within one year, although the work on the corpus will continue even after achieving this goal. On the other hand, we will continue testing different methods and tools for automating data collection from the corpus. The development of the EGPC will also refer to two main points of the use level: (1) the search tools that allow more granular searches and (2) the analysis tools that can structure extracted data according to different analysis criteria such as frequency, co-occurrence, word embedding, etc. This development sets up a possible move of the corpus to a new user environment.

One more direction in the development of the EGPC is adding new fields to it for other parallel corpora of Georgian with other languages. These corpora will be created and different bilingual projects will be implemented under the supervision and in cooperation with the Centre for Lexicography and Language Technologies at Ilia State University, including the framework of MA and PhD programs in lexicography at the University.

Thus our studies have revealed that parallel corpora are very useful tools for bilingual lexicography. Under-resourced languages like Georgian can balance lack of a large number of translated texts for parallel corpora by concentrating on the quality and data structure of the corpus and the lexical richness of text types and genres. It should be noted that balancing of a corpus concerns not only text genres (scientific, fiction, media), but also balanced amount of translations from a source language into a target language and vice versa. Such corpora can be conducive for compiling bilingual dictionaries, for enriching existing dictionaries with new terms, word meanings and illustrative collocations. Our study has also revealed the efficacy of high quality data of parallel

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1701 (Article)

sentences for machine translation, achieving positive results with much less data than are required by "resource-hungry" algorithms from the field of the NLP.

The methodology and the platform of a parallel corpus, created by our team, can also be used for the composition of parallel corpora in the languages other than English and Georgian.

Endnotes

1. https://opus.nlpl.eu/CCAligned/v1/en-ka_sample.html [Accessed 20.04.2022]
2. https://dumps.wikimedia.org/other/contenttranslation [Accessed 20.04.2022].
3. https://www.ted.com/participate/translate [Accessed 20.04.2022].
4. https://github.com/danielvarga/hunalign/2
5. https://terminotix.com/index.asp?name=SynchroTerm&content=item&brand=4&item=7&lang=en https://terminotix.com/index.asp?lang=en
6. The partner of Ilia State University in this project is Vakhtang Elerdashvili, a data scientist, a PhD Student at Text Technology Lab, Goethe-University Frankfurt, Germany, https://www.texttechnologylab.org/, the author of the Georgian spellchecker (https://spellchecker.ge/).
7. https://opennmt.net/
8. At present the testing of the program is underway in a closed intranet with the access only for the members of the working team.
9. https://www.google.com/search?q=google.translate+english+to+georgian&oq=google&aqs=chrome.2.69i60j46i67i131i199i433i465j35i39j69i60l4j69i65.4480j0j7&sourceid=chrome&ie=UTF-8 [Accessed 27.04.2022].

References

Dictionaries

[CEGD] ***Comprehensive English–Georgian Dictionary.*** Vol. I–XIV. (Editor-in-chief T. Margalitadze). 1995–2012. Tbilisi: Ivane Javakhishvili Tbilisi State University, Lexicographic Centre.

[CEGOD] ***Comprehensive English–Georgian Online Dictionary.*** (Editor-in-chief T. Margalitadze). 2010. Tbilisi: Ivane Javakhishvili Tbilisi State University, Lexicographic Centre.
Available at: www.dict.ge

[CGED] ***A Comprehensive Georgian–English Dictionary.*** 2 volumes. (Editor-in-chief D. Rayfield). 2006. London: Garnett Press.
Available at: http://www.nplg.gov.ge/gwdict/index.php?a=index&d=46

[EGPC] ***English–Georgian Parallel Corpus.*** Tbilisi: Ilia State University.
Available at: http://corp.dict.ge

Other references

Brezina, V. 2018. *Statistics in Corpus Linguistics: A Practical Guide.* Cambridge: Cambridge University Press.

Gippert, J. 2016. Complex Morphology and its Impact on Lexicology: The Kartvelian Case. Margalitadze, T. and G. Meladze (Eds.). 2016. *Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity, 6–10 September, 2016, Tbilisi, Georgia*: 16-36. Tbilisi: Ivane Javakhishvili Tbilisi State University.
Available at: http://euralex2016.tsu.ge/publication2016.pdf

Harris, A.C. (Ed.). 1991. *The Indigenous Languages of the Caucasus*: *Kartvelian. Vol. I.* Delmar, N.Y.: Caravan Press.

Kikvidze, Z. and L. Pachulia. 2019. Demetrius Rudolph Peacock and the Languages of Georgia. *General and Specialist Translation/Interpretation: Theory, Methods, Practice: International Conference Papers:* 15-22. Kyiv: Agrar Media Group.

Kubát, M. and J. Milička. 2013. Vocabulary Richness Measure in Genres. *Journal of Quantitative Linguistics* 20(4): 339-349.

Lison, P. and J. Tiedemann. 2016. OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. Calzolari, N. et al. (Eds.). 2016. *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), May 23–28, 2016, Portorož, Slovenia:* 923-929. Paris: European language Resources Association (ELRA).

Margalitadze, T. 2012. The Comprehensive English–Georgian Online Dictionary: Methods, Principles, Modern Technologies. Fjeld, R.V. and J.M. Torjusen (Eds.). 2012. *Proceedings of the 15th EURALEX International Congress, 7–11 August 2012, Oslo*: 764-770. Oslo: Department of Linguistics and Scandinavian Studies, University of Oslo.
Available at: http://euralex.org/category/publications/euralex-oslo-2012/

Margalitadze, T. 2014. European-Georgian Parallel Corpora for Georgian Lexicography and Translatology. *Proceedings of the International Conference 'Literary Translation — A Meeting Place for Nations and Literatures'.* Dedicated to the 100th Anniversary of a Translator, Poet and Theoretician of Literary Translation Givi Gachechiladze. Tbilisi: Ivane Javakhishvili Tbilisi State University.

Margalitadze, T. and G. Meladze. 2016. Importance of the Issue of Partial Equivalence for Bilingual Lexicography and Language Teaching. Margalitadze, T. and G. Meladze (Eds.). 2016. *Proceedings of the XVII EURALEX International Congress: Lexicography and Linguistic Diversity, Ivane Javakhishvili Tbilisi State University, Tbilisi, Georgia, 6–10 September, 2016*: 787-797. Tbilisi: Ivane Javakhishvili Tbilisi State University.
Available at: http://euralex.org/category/ publications/euralex-2016/

Margalitadze, T. and M. Odzeli. 2019. English–Georgian Dictionary *by Marjory Wardrop*. Tbilisi: Tbilisi State University Press.

Margalitadze, T. and Z. Pourtskhvanidze. 2019. The Georgian Language in AI-based Translation Models: Cooperation of Lexicographers and NLP Specialists. *EMLex Autumn Meeting and Colloquium, Tbilisi 2019, Georgia, October 8-11: Lexicography at a Crossroads.* Organized by TSU Lexicographic Centre and Consortium of European Master in Lexicography (EMLex).
Available at: https://margaliti.com/emlexweb.pdf

Margalitadze, T. and Z. Pourtskhvanidze. 2021. The Statistic-Based Mapping of the Distribution of Data Structure in a Parallel Corpus. International Conference *Languages in the Digital Age*, organized by State Language Department of Georgia, Centre for Language Technologies Tilde, under the patronage of the President of Georgia. October 2021.
Available at: http://enadep.gov.ge/uploads/Program_7_8_October_KA.pdf

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1701 (Article)

Margalitadze, T. and S. Tchighladze. 2022. *Unknown Pages of English–Georgian Lexicography*. Tbilisi: Ilia State University Press.

Reimers, N. and I. Gurevych. 2020. Making Monolingual Sentence Embeddings Multilingual Using Knowledge Distillation. *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP):* 4512-4525. Online: Association for Computational Linguistics.

Tiedemann, J. 2012. Parallel Data, Tools and Interfaces in OPUS. Calzolari, N. et al. (Eds.). 2012. *Proceedings of the Eighth International Conference on Language Resources and Evaluation, Istanbul, Turkey, May 21–27, 2012* (LREC 2012): 2214-2218. Istanbul, Turkey: European Language Resources Association (ELRA).

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1702 (Article)

Towards a Comprehensive Dictionary of Gabonese French

Paul A. Mavoungou, *Département des Sciences du Langage, Université Omar Bongo, Libreville, Gabon; Centre de Recherche en Etudes Germaniques et Interculturelles, Université Omar Bongo and Department of Language Education, University of the Western Cape, South Africa (moudika2@yahoo.fr)*

Hugues Steve Ndinga-Koumba-Binza, *Department of Language Education, University of the Western Cape, Bellville, South Africa (nkbinza@uwc.ac.za)*

Virginie Ompoussa, *Département des Sciences de l'Information et de la Communication, Université Omar Bongo; Département des Sciences du Langage, Université Omar Bongo, Libreville, Gabon and Groupe de Recherche en Langues et Cultures Orales, Université Omar Bongo (vompoussa@yahoo.com)*
and
Blanche Nyingone Assam, *Department of Foreign Languages, University of the Western Cape, Bellville, South Africa (bassam@uwc.ac.za)*

Abstract: The present article reports on the conceptualization of the *Dictionnaire Général du Français Gabonais*. The dictionary project is a first of its kind in Gabonese lexicography. As an outcome of the inception of Gabonese French lexicography, the dictionary project arose from a discussion on the definition of Gabonese French, which Gabonese French lexicography should account for. In this article, the project background as well as the interests for the planned dictionary are presented. The article also deals with two key aspects of the dictionary conceptualization plan, i.e., lexicographic processes and the dictionary basis.

Keywords: CONCEPTUALIZATION PLAN, DICTIONARY BASIS, DICTIONARY PROJECT, FRENCH, GABON, LANGUAGE APPROPRIATION, LEXICOGRAPHIC PROCESSES

Résumé: Vers un Dictionnaire Général du Français Gabonais. Le présent article rend compte de la conceptualisation du *Dictionnaire Général du Français Gabonais*. Le projet de ce dictionnaire est une première du genre dans la lexicographie gabonaise. Issu du lancement de la lexicographie française gabonaise, le projet de ce dictionnaire est né d'une réflexion sur la définition du français gabonais, dont la lexicographie française gabonaise doit tenir compte. Dans cet article, le contexte du projet ainsi que les intérêts pour le dictionnaire envisagé sont présentés.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1702 (Article)

L'article traite également de deux aspects clés du plan de conceptualisation du dictionnaire, à savoir les processus lexicographiques et la base du dictionnaire.

Mots-clés: APPROPRIATION LINGUISTIQUE, BASE DU DICTIONNAIRE, FRANÇAIS, GABON, PLAN DE CONCEPTUALISATION, PROCESSUS LEXICOGRAPHIQUES, PROJET DE DICTIONNAIRE

1. Introduction

Responding to Mavoungou's (2013) definition of Gabonese French and discussion of Gabonese French lexicography, Nyangone Assam et al. (2016) lay out among other issues the position of Gabonese French within the emerging Gabonese lexicography, "especially in terms of metalexicographical research" (Nyangone Assam et al. 2016: 184). As a response to the "strategic focus on Gabonese French lexicography" (Nyangone Assam et al. 2016: 182-183), the project for the planning of a monolingual *Dictionnaire Général du Français Gabonais* (henceforth DGFG), i.e., a monolingual comprehensive dictionary of Gabonese French, was launched by the *Centre de Recherche en Etudes Germaniques et Interculturelles* (CREGI) at Omar Bongo University in Libreville, Gabon. The aim of the projected dictionary is to present the full spectrum of the lexicon of Gabonese French.

The present article reports on the planning and compilation of the DGFG. It focuses on a few matters that relate to the conceptualization plan of the projected dictionary. In general, the purpose of the study is to contribute to the grounding of Gabonese French lexicography within the strategic planning of Gabonese lexicography. However, more specifically, the aim of this article is twofold. First, it seeks to present the conceptualization principles of the intended dictionary. Second, it is a follow-up to the recent debate on the content of Gabonese French dictionaries. The article starts in section 2 with an overview of the background of the dictionary project. Section 3 presents the target users and the interests of the planned dictionary, while section 4 deals with the lexicographic processes. In section 5, the dictionary basis of the projected dictionary is discussed.

2. Background of the dictionary project

The background of the intended dictionary is primarily found in the theoretical inception of Gabonese French lexicography, the debate that took a stand on the definition of Gabonese French and the contents of its dictionaries, and the language appropriation that the Gabonese have made of the French language. The outline of the background of this dictionary is therefore a response to Nyangone Assam et al. (2016), who argue to produce Gabonese French dictionaries based on an accurate understanding of the concept of Gabonese French.

The article by Nyangone Assam et al. (2016), in response to Mavoungou (2013) on the definition of the concept of Gabonese French, laid out theoretical grounds for the inception of Gabonese French lexicography.

2.1 Origins and theoretical inception of Gabonese French lexicography

Mavoungou (2013) and Nyangone Assam et al. (2016) have been regarded as the first conceptualizations of Gabonese French lexicography, i.e., "the set of theoretical and practical works done on the French language as it is spoken in Gabon" (Nyangone Assam et al. 2016: 178). However, it is important to note that the theoretical inception of Gabonese French lexicography started with a book chapter by Mavoungou (2002), who suggested a dictionary of French as spoken in Gabon. Based on the suggestions by Mavoungou (2002), a short dictionary manuscript followed, containing a hundred dictionary articles (Mavoungou et al. 2002). In addition, Mavoungou (2011), Nsa Ndo (2010) and Nsafou (2010) proposed that various forms of lexical particularisms of French as spoken in Gabon be identified and recommended their lexicographical treatment within a monolingual dictionary.

One of the major shortcomings of the proposals by Mavoungou (2002, 2011), Mavoungou et al. (2002), Nsa Ndo (2010) and Nsafou (2010) is the exclusion of the acrolect variant of the language, i.e., the standard variety known as Parisian French. As highlighted in Nyangone Assam et al. (2016), the variety of the French language as spoken in Gabon has been recognized not only as one of the varieties of this worldwide language, but also as one of the local languages of naturally and culturally multilingual but officially monolingual Gabon. In the field of lexicography, attention has particularly been given to the lower mesolectal forms (popular Gabonese French), basilectal forms (Gabonese Matitis French) as well as to slang forms (Toli-bangando) with the production of a series of dictionaries (Boucher and Lafage 2000; Ditougou 2009; Dodo-Bounguendza 2008, 2010, 2013; Moussounda Ibouanga 2011; and Mavoungou et al. 2014, 2015).

The discussion on the content of the previous Gabonese French dictionaries has reached the conclusion that these do not represent the full scope of Gabonese French. As a core component of the emerging Gabonese lexicography, Gabonese French lexicography is expected to produce Gabonese French dictionaries, which may locally replace imported French dictionaries compiled in France. It should be noted that several metalexicographical studies have also been conducted on Gabonese French. These studies can be divided into three groups. The first group contains the studies that initiated the inception of Gabonese French lexicography (Mavoungou et al. 2002; Mavoungou 2002; Nyangone Assam et al. 2016). The second group of studies is that of the planning of Gabonese French dictionaries (Nsa Ndo 2010; Nsafou 2010; Mouélé 2011). The third group comprises works that critique published dictionaries of Gabonese French (Mavoungou 2002, 2011; Ondo Mébiame and Ekwa Ebanéga 2011; Nyangone Assam et al. 2016).

Unfortunately, except for Nyangone Assam et al. (2016) who suggested a strategic focus including "metalexicographic research and corpus building for all types of monolingual dictionaries of Gabonese French" (Nyangone Assam et al. 2016: 184), none of the previous studies makes a case for a comprehensive dictionary of Gabonese French. The projected *Dictionnaire Général du Français Gabonais* is set to avoid the flaws of the previous Gabonese French dictionaries. The "strategic focus on Gabonese French lexicography" as suggested by Nyangone Assam et al. (2016: 182-183) clearly advocates monolingual lexicography and monolingual dictionary production as far as Gabonese French is concerned. This principle has been adopted within CREGI for the launching of Gabonese French lexicography through the planned *Dictionnaire Général du Français Gabonais*. CREGI also intends to initiate research for a *Corpus de Français Gabonais* (CFG), i.e., a corpus of Gabonese French on the model of *Corpus de la Langue Française* (André 2017; Siepmann et al. 2016; Equipe DELIC 2004), to which the CFG should contribute.

2.2 Gabonese French: Clarifications towards concluding the debate

Mavoungou (2013: 260) defined Gabonese French as "a repertoire of a variety of lexical items and expressions". Nyangone Assam et al. (2016: 167) refuted this definition as being reductive. This reductive definition is also evident in Mabika Mbokou's (2019: 2) claim that "to speak French in Gabon is to speak a French whose lexicon, meaning of words and their use is different from the norm of standard French"[1].

As Nyangone Assam et al. (2016: 167) put it, "the French language that is the official language of the Republic of Gabon (according to its Constitution), the second language of the current political, intellectual and administrative elites of Gabon, and the mother tongue of the majority of the Gabonese youth cannot be a form of language made only of lexical, phrasal and pronunciation particularisms".

Furthermore, Nyangone Assam et al. (2016: 176) also refute the stratification of the varieties of Gabonese French suggested by Mavoungou (2013: 259). In fact, according to Mavoungou (2013: 259), four distinct language varieties can be found in Gabonese French, i.e., Standard French, Official French (Acrolectal level), Common French (Mesolectal level) and Popular French (Basilectal level). Nyangone Assam et al. (2016: 172) however believe that the term "official French" cannot be identified with a sociolect. Moreover, it appears possible to make a distinction between upper mesolect and lower mesolect within Gabonese French.

The compilation of the projected DGFG is intended to be based on the conception of Gabonese French schematized as follows by Nyangone Assam et al. (2016: 176) in Figure 1 below. This schema is an actual stratification with all identified sociolectal variants of Gabonese French.

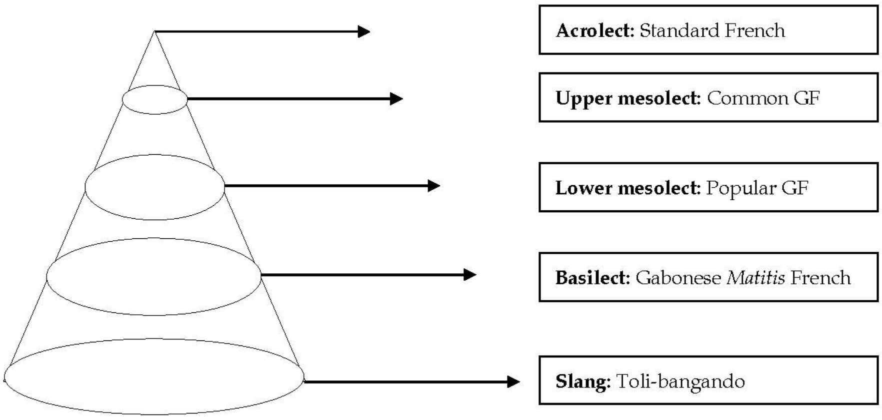

Acrolect: Standard French
Upper mesolect: Common GF
Lower mesolect: Popular GF
Basilect: Gabonese *Matitis* French
Slang: Toli-bangando

Figure 1: Sociolectal stratification of Gabonese French (Nyangone Assam et al. 2016: 176)

Subsequently, Nyangone Assam et al. (2016: 176) also reject the distinction made by Mavoungou (2013) between variant A and variant B where variant A is supposedly the variety of France and variant B the Gabonese French. In our view, "each French variety in a given country has to be legitimized, i.e., accepted a one of the various speech-forms of a particular language following the codification of such a variant in a determined country" (Nyangone Assam et al. 2016: 176-177).

In the light of this understanding, what do we really mean by Gabonese French? Gabonese French must be understood as the total set of phonetic, phonological, morphological, syntactical, and lexical features of the French language as it is spoken in Gabon. Finally, it can be said that Gabonese French is one of the varieties of French (a worldwide language), but also one of the local languages of culturally multilingual but officially monolingual Gabon.

2.3 *Français du Gabon* or *Français Gabonais*? A language appropriation approach

The above understanding of the concept of Gabonese French may also decide on its naming in French. In fact, French-language literature about Gabonese French often designates the latter in the following three ways (from the most to the least frequently occurring):

(i) *Français au Gabon*, i.e. French in Gabon (Massinga Kombila 2013; Mindze M'Eyeghe 2001; Pambou 1998; Boucher 1997; Artigues 1995; Boutin-Dousset 1989; Moussirou Mouyama 1984),

(ii) *Français du Gabon*, i.e. Gabon's French (Mabika Mbokou 2019; Italia 2006; Mavoungou 2002; Mouloungui Nguimbyt 2002), or

(iii) *Français Gabonais*, i.e. Gabonese French (Mouélé 2011; Minko 2008; Mitchell 2004)

The semantics of these three terms depicts the following perceptions of the French language in Gabon:

(i) "French in Gabon" depicts a foreign language locally used.

(ii) "Gabon's French" refers to a sense of origins (from) or of possession (of). Is it a language that comes from Gabon or a language that is a propriety of Gabon? It is a fact that French does not come from Gabon but can be considered as a property of Gabon only on the grounds that the language is recognized as the sole official language by the Constitution of the Republic of Gabon.

(iii) "Gabonese French" depicts language appropriation as it has been reflected in several studies (Pambou 1998; Mabika Mbokou 2008, 2012, 2019; Ndinga-Koumba-Binza 2011; Massinga Kombila 2013; Boussougou and Menacere 2015).

The compilation of the projected DGFG assumes French to be a Gabonese language. It is the same sense of appropriation that transpires in the term "South African English" (*anglais sud-africain*), which may have a whole different meaning than "English in South Africa" (*anglais en Afrique du Sud*) or "South Africa's English" (*anglais d'Afrique du Sud*). The same can be said about concepts such as "Canadian French" (*français canadien*) as compared to "French in Canada" (*français au Canada*) and to "Canada's French" (*français du Canada*); and about "Belgian French" (*français belge*) compared to "French in Belgium" (*français en Belgique*) and "Belgium's French" (*français de Belgique*).

It should be noted that the terms "Canadian French" (Walker 1984; Martineau 2007; Poder et al. 2021; Attieh et al. 2022), "Belgian French" (Hambye and Simon 2012; Pedraza and Cougnon 2021) and Swiss French (Sertling Miller 2007; Racine and Andreassen 2012) are fully accepted terms in the literature in English as well as in French (Mougeon and Beniak 1989; Martineau 2005; Andreassen 2018; Pigeon 2021). About Canadian French, it may be worth mentioning the statement by Poliquin (2006: 4), a French-speaking Ontarian and linguist: "I have grown up referring to my dialect as 'français canadien' not 'français québécois'".

In the general context of Africa, Zabus (2007) views this foreign language appropriation as a form of indigenization. According to Zabus (2007), who demonstrates the indigenization process of both English and French through novels and other works of fiction, the indigenization is both in the text and in the context (Zabus 2007: 4-8). In terms of the context, Mengara (2000: 282) argues

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1702 (Article)

that French "has become an African language". According to Mengara (2000: 282-283), linguistic research "has yielded results which have all pointed towards the fact that the French used in Africa now presents a certain number of structural differences from the French of France — differences that can be seen at the lexical, syntactic, semantic, phonological and cultural levels, as a result of the variety of linguistic loans, extensions, calques, reductions, transfers, etc."

This type of phenomenon is always likely to occur when a language extends beyond its original borders, and notably in a context of political, economic and cultural imperialism. This is how regional dialects are born. For French in Africa, Mengara (2000: 283) argues that these "differences were significant enough for researchers to start talking about a different code or dialect of French that they called African French". The reality is that the so-called African French is an aggregate of numerous varieties spoken in various parts of Africa, depending on the level and type of appropriation and codification each region or each country has made of the French language. This explains why besides Gabonese French we can talk about Cameroonian French (Eloundou Eloundou 2019; Djoum Nkwescheu 2008), Ivorian French (Atsé N'Cho 2018; Plahar 2017; Kouadio N'Guessan 2008), Malian French (Diarra 2018; Skattum 2010), Moroccan French (Nifaoui 2021; Benzakour 2010), Senegalese French (Fall 2021; N'Diaye Corréard 2008), etc. All of this speaks to the appropriation of the French language in the different African countries. This appropriation can be referred to as a form of "decolonizing French as a foreign Language" (Nel and Ferreira-Meyers 2020: 1).

The appropriation of the French language as understood in the projected DGFG is in line with the conception of the Gabonese language landscape (Ndinga-Koumba-Binza 2005, 2007) in which French is viewed as a local language rather than a foreign language. At the same time, the constitutional disposition making French as the sole official language of Gabon is the legal ground for the nationalisation of the French language in Gabon as a Gabonese language (Ndinga-Koumba-Binza 2011, 2013). As such, the term "Gabonese French" (*français gabonais*) is believed to be more appropriate to designate the language as it is spoken in Gabon.

3. Research interest and target users

Both lexicographic and metalexicographic works (Mavoungou 2002, 2013; Mavoungou et al. 2014, 2015; Nyangone Assam et al. 2016) have shown the necessity for Gabonese lexicographers to come up with a monolingual comprehensive dictionary of Gabonese French. This posits the research interest of the current project. It is a well-known fact that to date the only French dictionaries used in Gabon are the dictionaries compiled and produced in France.

Moreover, in all existing publications dedicated to the French language as it is spoken in Gabon, the historical dimension is not systematically considered. For the purpose of the projected dictionary, the historical dimension is under-

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1702 (Article)

stood as the etymology for a given lemma as well as the different meanings of a particular lexical item, starting from the most recent sense to the oldest one.

Considering such a research interest, the question of the specific target users of the DGFG does arise. The target users of the planned dictionary are the Gabonese public at large. This group of envisaged users is basically composed of second-language speakers of Gabonese French as well as first-language speakers of Gabonese French. It is a very heterogeneous group which may have different lexicographic needs and reference skills that should be catered for in the dictionary design. The speakers of the French language in Gabon are indeed very diverse. Pambou (1998) and Ndinga-Koumba-Binza (2011) highlight the fact that the language has multiple statuses in Gabon. Being the sole official language of the country, there are speakers who use it as a foreign language, some as a second language and some other as their mother tongue. Apart from the 10 000 and plus French citizens who have settled in Gabon (Ndinga-Koumba-Binza 2011: 138), Mabika Mbokou (2012: 172) indicates that the majority of the younger Gabonese generations, especially those in urban areas, have French as their initial language or mother tongue. The existence of a group of target users for the DGFG is therefore a certainty.

To add on the heterogeneity of the target users' group in which many ethnolinguistic groups can be found, there are some common traits among Gabonese, namely the worldview, the value system, traditional beliefs, native languages, and cultures, etc. Among Gabonese, migrants and foreigners (from Europe, Africa, America, Asia, etc.) living in Gabon and knowing French, there might be a lot of differences regarding their ways of life, traditions, common beliefs, institutions and collective activities.

The planned dictionary should assist both second and first-language speakers of Gabonese French (pupils, students, civil servants, politicians, journalists, business people, missionaries, militaries, etc.). To be successful, the publication of a given dictionary project must be the result of proper planning (Gouws 2001: 64). The following sections present the main features of the dictionary conceptualization plan of DGFG.

4. Comprehensive lexicographic processes

Gouws (2001) agrees with Wiegand (1998) that the dictionary conceptualization plan can be divided into five phases, namely:

(i) the general preparation phase,

(ii) the material acquisition phase,

(iii) the material preparation phase,

(iv) the material processing phase, and

(v) the publishing preparation phase.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1702 (Article)

The dictionary conceptualization plan of a given lexicographical project forms part of a lexicographical process. Gouws and Prinsloo (2005: 9) indicate that a lexicographic process is a "comprehensive set of activities" of which the compilation of the dictionary is the culminating result. A lexicographical process refers to all the activities leading to the publication of a given project. It is "part of a comprehensive historical process which coincides with the development of a language" (Gouws 2001: 65).

A distinction is usually made between primary and secondary lexicographical processes. The primary comprehensive lexicographical process (Gouws 1999: 7-10) refers to all the decisions taken at State or Government level in order to plan, promote, guide, and develop lexicographical activities (language policy issues, the establishment of National Lexicographical Units or NLUs, the training of staff members for a particular NLU or a specific dictionary project, the use of a unified orthography for all the NLUs, etc.).

The decision to declare French as the sole official language of Gabon was arguably in itself the commencement of the lexicographic process towards a dictionary of Gabonese French. This has never been followed by any other official recognition of French as an item of the local linguistic heritage of Gabon. Remarkably, the government of Gabon has, since attaining independence, never made any public decision in favour of the French language other than the constitutional provision. Nevertheless, the late President Omar Bongo repeatedly voiced his opinion that the French language, being the only tool for interethnic communication, was the foundation of national unity in the country (Bongo Ondimba 1998) and publicly referred to French as "our national language" (Ndinga-Koumba-Binza 2011: 146). Ultimately, in terms of planning at a macro level, the compilation of the DGFG is the result of these unconfirmed government processes with regard to the French language in Gabon.

The secondary comprehensive lexicographical process refers to all the activities conducted within a given lexicographical unit at national level or within a particular dictionary project. These activities may include editorial work or issues such as the formulation and implementation of the organisation plan of a given dictionary project, the identification of short-, medium- and long-term objectives within a project, liaisons with publishers, the planning of the marketing of each project, the logistics of the project and all the managerial aspects (Gouws 2001: 65).

It should be noted that in Gabon no lexicographical unit has been set at national level despite numerous pleas from Gabonese lexicographers (Emejulu 2000, 2001, 2002, 2003; Ndinga-Koumba-Binza 2005; Mavoungou 2010). Three noteworthy initiatives in this regard have however been put in place at academic level. The first is the adoption by GRELACO[2] of a programme for the development of lexicography as a research discipline and a career (Emejulu 2000, 2001) that can contribute to the promotion of Gabonese native languages. This programme, which is still running, has resulted in the training of Gabonese lexicographers at doctoral level, the development of lexicography as a teaching

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1702 (Article)

and research discipline, the inception of Gabonese lexicography and the advent of a modern era of Gabonese language dictionary production (Ndinga-Koumba-Binza 2006). The projected DGFG can be viewed as the result of this lexicographic process initiated by GRELACO in 1998.

A second initiative is the recently established GREDYLEX (*Groupe de Recherche sur les Dynamiques Linguistiques et Lexicographiques*), a research unit within the Institute for Research in Human Sciences (IRSH) of the National Centre for Science and Technology Research (CENAREST). GREDYLEX is the first of its kind within IRSH and CENAREST, following numerous attempts to boost linguistic and lexicographic research. The mission of GREDYLEX can be clearly summarized as the development of lexicography practice and a dictionary culture within Gabon. It is therefore no surprise that the research and community engagement activities of GREDYLEX are mainly centered on pedagogical lexicography and schools. This is clearly reflected in their three periodicals, namely *Kabi*, a journal for practical lexicographers; *Likayi*, their newsletter and academic information bulletin; and *Ilongo*, a magazine for information aimed at the general public. The first volumes of the three publications were released soon after the launching of GREDYLEX at the end of 2021.

Understanding that French is the sole language of learning and teaching (LoLT) in Gabon's education system, research at GREDYLEX has a strong focus on French and the use of French dictionaries encyclopedias and textbooks. The consideration of the French language alongside the Gabonese native languages is evident in the publications and mission of GREDYLEX. This process is taken into consideration in the planned DGFG. The editorial team and the compilers of the projected DGFG intend to collaborate with GREDYLEX in various aspects of the production of the DGFG.

Finally, as initially mentioned, the DGFG project originates from the CREGI (*Centre de Recherche en Etudes Germaniques et Interculturelles*) within the Department of German Studies at Omar Bongo University. CREGI was launched in January 2013 with the aim to serve as the research wing of the Department of German Studies. The fact that the Department has four lexicographers (two permanent and two associates) who are also founding members of CREGI contributed to retaining lexicography and lexicology as one of the research areas of CREGI's activities. In addition to the annual dictionary colloquium, the DGFG project is one of CREGI's main lexicographical activities and projects.

In our view GRELACO, GREDYLEX and CREGI constitute potential springboards or starting points for the establishment of national lexicography units (NLUs), based on the model of the South African NLUs. While the contribution and the role of NLUs in South Africa have been analyzed (Kumalo 1999; Gouws 2003; Mongwe 2006), Madiba and Nkomo (2010: 322) believe that "The establishment of the NLUs remains a commendable idea which has undoubtedly improved lexicographic practice in the country". Within the DGFG project, the South African NLUs model is recommended. What may be missing is an

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1702 (Article)

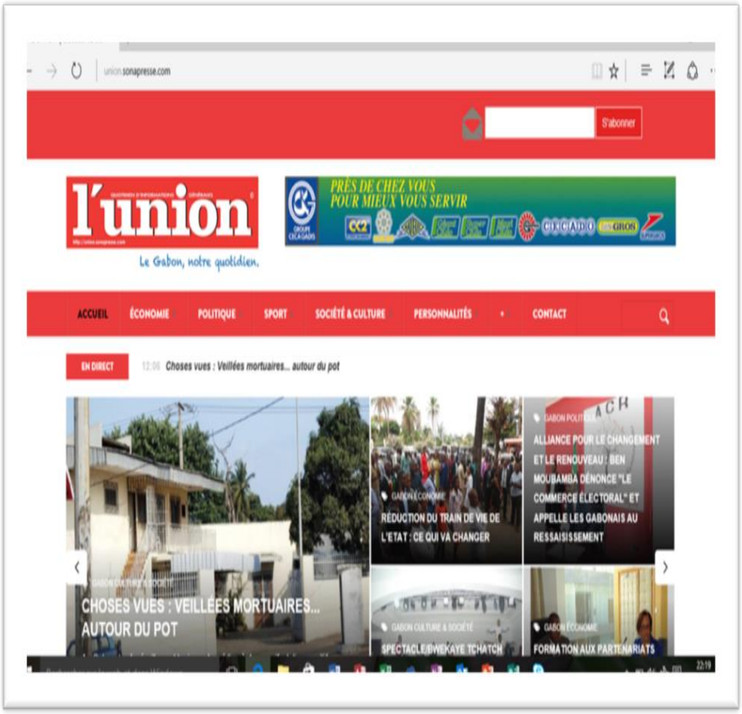
organization such as the Pan South African Language Board (PanSALB) to coordinate between the NLUs and provide government tutelage.

5. The dictionary basis of DGFG

A dictionary basis refers to the set of sources "used for the compilation of a dictionary" (Svensén 2009: 39). It can also be said that the dictionary basis is the total of the source language material for a specific lexicographic project. Three types of sources can be identified, namely *primary sources* (both written material and oral sources), *secondary sources* (all available dictionaries in the specific language) and *tertiary sources* (all other linguistic material such as grammars, scientific articles, and books).

5.1 Primary sources: written materials

As far as primary sources are concerned, the written material of DGFG will be extracted from various newspapers published in Gabon. The major source amongst these publications will be the sole national daily newspaper, namely *L'Union*. Figure 2 below presents a screenshot from the website of this media.



12:06 Choses vues : Veillées mortuaires... autour du pot

GABON CULTURE & SOCIÉTÉ
CHOSES VUES : VEILLÉES MORTUAIRES... AUTOUR DU POT

GABON ÉCONOMIE
RÉDUCTION DU TRAIN DE VIE DE L'ETAT : CE QUI VA CHANGER

GABON POLITIQUE
ALLIANCE POUR LE CHANGEMENT ET LE RENOUVEAU : BEN MOUBAMBA DÉNONCE "LE COMMERCE ÉLECTORAL" ET APPELLE LES GABONAIS AU RESSAISISSEMENT

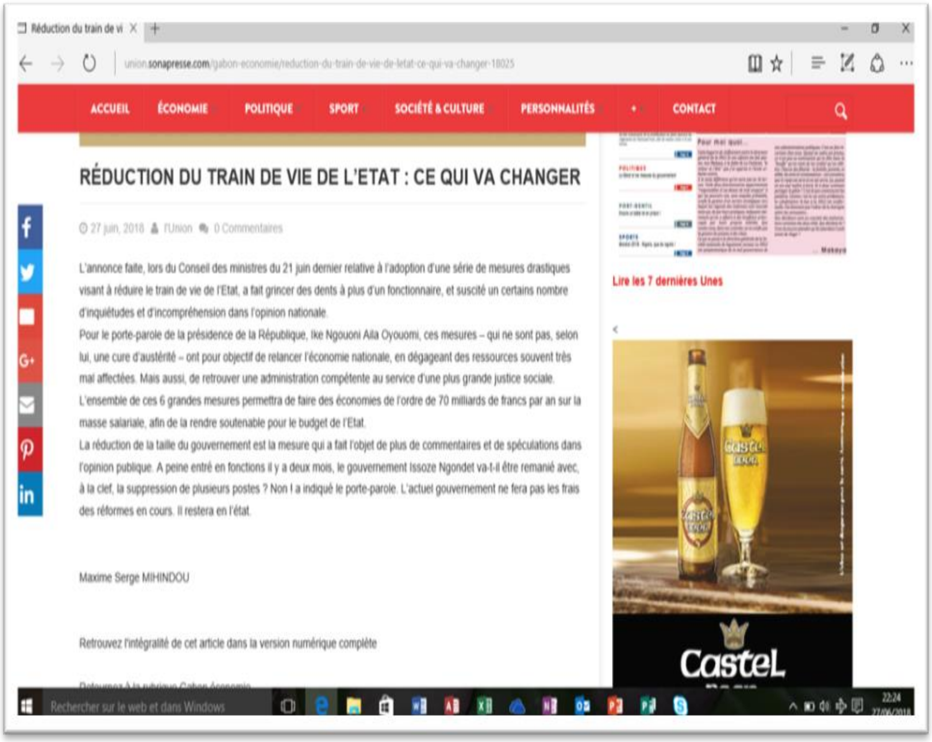
GABON CULTURE & SOCIÉTÉ
SPECTACLE/BWEKAYE TCHATCH

GABON ÉCONOMIE
FORMATION AUX PARTENARIATS

Figure 2: Homepage of *L'Union* website[3]

As in most newspapers for the public at large, articles in *L'Union* are broad in topics and very explanatory. The online version of the newspaper is the exact copy of the printed version in terms of content, structure, and length of articles. This provides an available source for electronic texts for lexicographic works such as corpus building and dictionary compilation.

Figure 3 below gives an illustration of an article in the online version of *L'Union*.


Réduction du train de vi
union.sonapresse.com/gabon-economie/reduction-du-train-de-vie-de-letat-ce-qui-va-changer-18025
ACCUEIL
ÉCONOMIE
POLITIQUE
SPORT
SOCIÉTÉ & CULTURE
PERSONNALITÉS
+
CONTACT
Lire les 7 dernières Unes
Castel
Rechercher sur le web et dans Windows
22:24
27/06/2018

RÉDUCTION DU TRAIN DE VIE DE L'ETAT : CE QUI VA CHANGER

27 juin, 2018 l'Union 0 Commentaires

L'annonce faite, lors du Conseil des ministres du 21 juin dernier relative à l'adoption d'une série de mesures drastiques visant à réduire le train de vie de l'Etat, a fait grincer des dents à plus d'un fonctionnaire, et suscité un certains nombre d'inquiétudes et d'incompréhension dans l'opinion nationale.

Pour le porte-parole de la présidence de la République, Ike Ngouoni Aila Oyouomi, ces mesures – qui ne sont pas, selon lui, une cure d'austérité – ont pour objectif de relancer l'économie nationale, en dégageant des ressources souvent très mal affectées. Mais aussi, de retrouver une administration compétente au service d'une plus grande justice sociale.

L'ensemble de ces 6 grandes mesures permettra de faire des économies de l'ordre de 70 milliards de francs par an sur la masse salariale, afin de la rendre soutenable pour le budget de l'Etat.

La réduction de la taille du gouvernement est la mesure qui a fait l'objet de plus de commentaires et de spéculations dans l'opinion publique. A peine entré en fonctions il y a deux mois, le gouvernement Issoze Ngondet va-t-il être remanié avec, à la clef, la suppression de plusieurs postes ? Non ! a indiqué le porte-parole. L'actuel gouvernement ne fera pas les frais des réformes en cours. Il restera en l'état.

Maxime Serge MIHINDOU

Retrouvez l'intégralité de cet article dans la version numérique complète

Figure 3: Extract from *L'Union* website[4]

Apart from *L'Union*, Gabon's first and biggest newspaper, written data can also be obtained from the weekly *Gabon d'Aujourdhui*, which is published by the Ministry of Communications, as well as from other periodicals. There are several privately owned periodicals, which are either independent or affiliated with political parties.

Table 1 below provides a list of Gabonese online media and Gabonese newspapers that use Gabonese French. A number of these newspapers are only online publications. Most of the paper newspapers are also published numerically either through their own respective websites or on the e-kiosk of their common distributor[5].

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1702 (Article)

Name	Language	Frequency	Ownership	Notes
7jours Infos	French	Daily	Private	Online
Coopération Internationale	French	Monthly	Private	Paper & digitized
Dépêches241	French	Daily	Private	Online
Echos du Nord	French	Weekly	Private	Paper & digitized
Ezombolo	French	Bi-monthly	Private	Paper & digitized
Gabon Actu	French	Daily	Private	Online
Gabon d'Aujourdhui	French	Weekly	Government	Paper
Gabon Eco	French	Daily	Private	Online
Gabon Libre	French	Daily	Private	Online
Gabon Matin	French	Daily	Government	Paper
Gabon Media Time	French	Daily	Private	Online
Gabon Review	French	Daily	Private	Online
Gabonews	French	Daily	Private	Online
Infos Plus Gabon	French	Daily	Private	Online
Infos241	French	Daily	Private	Online
L'Aube	French	Weekly	Private	Paper & digitized
L'Union	French	Daily	Government	Paper & digitized
La Calotte	French	Bi-monthly	Private	Paper & digitized
La Cigale Enchantée	French	Weekly	Private	Paper & digitized
La Loupe	French	Weekly	Private	Paper & digitized
La Lowe	French	Weekly	Private	Paper
La Nation	French	Bi-monthly	Private	Paper & digitized
La Nouvelle Republique	French	Bi-monthly	Private	Paper & digitized
La Relance	French	Weekly	Private	Paper
Le Mbandja	French	Weekly	Private	Paper & digitized
Le Temoin	French	Weekly	Private	Paper
Le Temps	French	Weekly	Private	Paper
Moutouki	French	Weekly	Private	Paper & digitized
Tango	French	Bi-monthly	Private	Paper & digitized

Table 1: List of the most important periodicals using Gabonese French[6]

These media, especially those that are independent, are full of cartoons and therefore include a wealth of words and expressions representing lower mesolectal forms (popular Gabonese French), basilectal forms (Gabonese Matitis French) as well as slang forms (Toli-bangando). As such, these cartoons are actual linguistic and cultural containers of knowledge (cf. McArthur 2006).

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1702 (Article)

The works of well-known Gabonese cartoonists will also contribute to the dictionary basis for both texts and illustrative pictures — cf. the cartoons by Patrick Essono (Pahé) (Figure 4) and Landry Békalé (Lybek) (Figure 5).

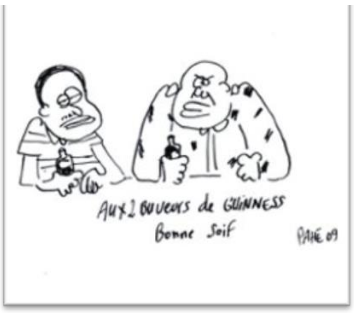


Figure 4: Extract of Pahé in the online newspaper *Gaboneco*[7]

The decision to consider the work of cartoonists in the dictionary basis is based on three considerations. First, it portrays the Gabonese society on a daily basis with a particularly biting/sharp sense of humour. Figure 5 below shows the heading "Gabonitudes" by Lybek in *L'Union*. It is believed that "Gabonitudes" is the most consulted text of the newspaper.

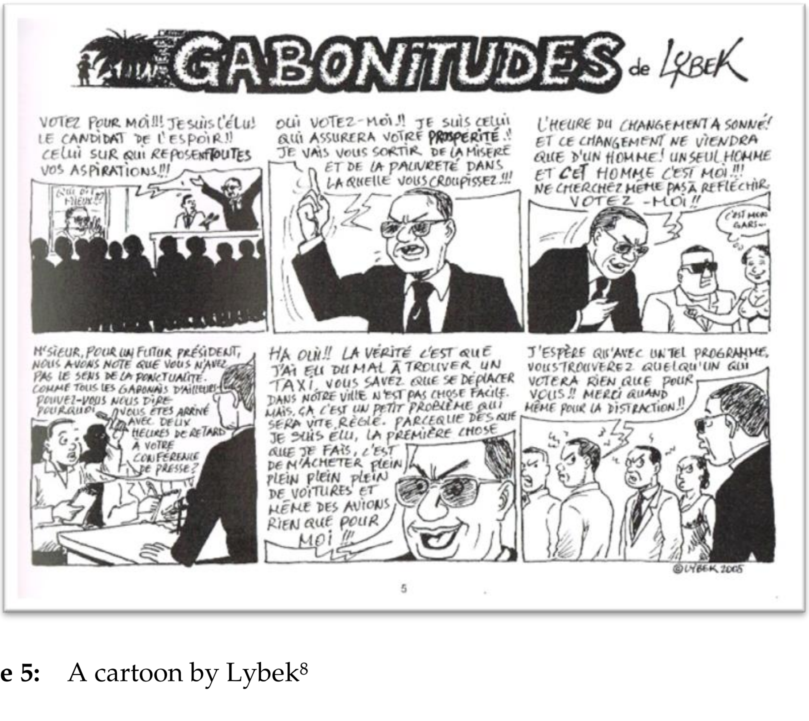

GABONITUDES de LYBEK
VOTEZ POUR MOi!!! JE SUiS L'ÉLU! LE CANDiDAT DE L'ESPOiR!! CELUi SUR QUi REPOSENT TOUTES VOS ASPiRATIONS!!!
QUi DiT MIEUX!?
OUi VOTEZ-MOi!! JE SUiS CELUi QUi ASSURERA VOTRE PROSPERiTÉ!! JE VAiS VOUS SORTiR DE LA MiSÈRE ET DE LA PAUVRETÉ DANS LAQUELLE VOUS CROUPISSEZ!!!
L'HEURE DU CHANGEMENT A SONNÉ! ET CE CHANGEMENT NE ViENDRA QUE D'UN HOMME! UN SEUL HOMME ET CET HOMME C'EST MOi!!! NE CHERCHEZ MÊME PAS À REFLÉCHiR. VOTEZ-MOi!!
C'EST MON GARS…
M'SiEUR, POUR UN FUTUR PRÉSIDENT, NOUS AVONS NOTÉ QUE VOUS N'AVEZ PAS LE SENS DE LA PONCTUALiTÉ. COMME TOUS LES GABONAiS D'AiLLEURS! POUVEZ-VOUS NOUS DiRE POURQUOi VOUS ÊTES ARRiVÉ AVEC DEUX HEURES DE RETARD À VOTRE CONFÉRENCE DE PRESSE?
HA OUi!! LA VÉRiTÉ C'EST QUE J'Ai EU DU MAL À TROUVER UN TAXi. VOUS SAVEZ QUE SE DÉPLACER DANS NÔTRE ViLLE N'EST PAS CHOSE FACiLE. MAiS, ÇA C'EST UN PETiT PROBLÈME QUi SERA ViTE RÉGLÉ. PARCEQUE DÈS QUE JE SUiS ÉLU, LA PREMiÈRE CHOSE QUE JE FAiS, C'EST DE M'ACHETER PLEiN PLEiN PLEiN PLEiN DE VOiTURES ET MÊME DES AViONS RiEN QUE POUR MOi!!!
J'ESPÈRE QU'AVEC UN TEL PROGRAMME, VOUS TROUVEREZ QUELQU'UN QUi VOTERA RiEN QUE POUR VOUS!! MERCi QUAND MÊME POUR LA DiSTRACTiON!!
©LYBEK 2005
5

Figure 5: A cartoon by Lybek[8]

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1702 (Article)

The best of *Gabonitudes* has been published in a comic book entitled *Gabonitudes Tome I.* Figure 6 below shows the front cover.

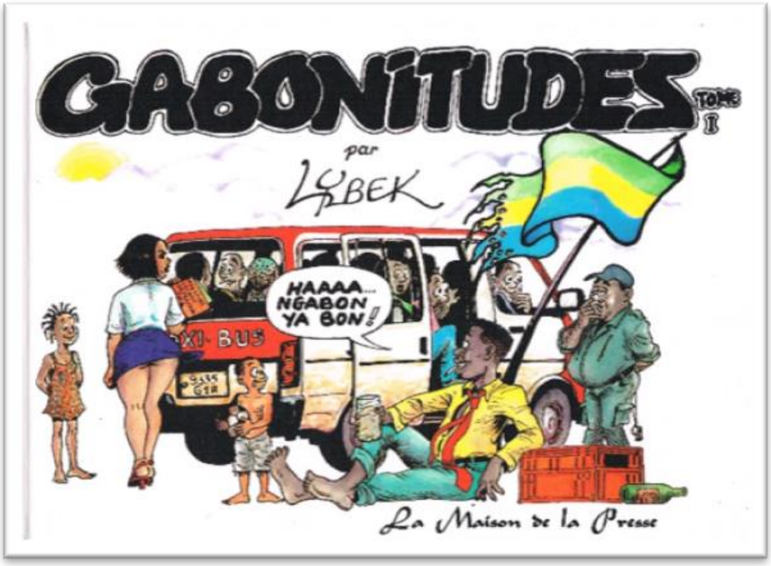


Figure 6: The front cover of *Gabonitudes Tome I*

Figure 7 below presents the back cover of the same comic book by Lybek.

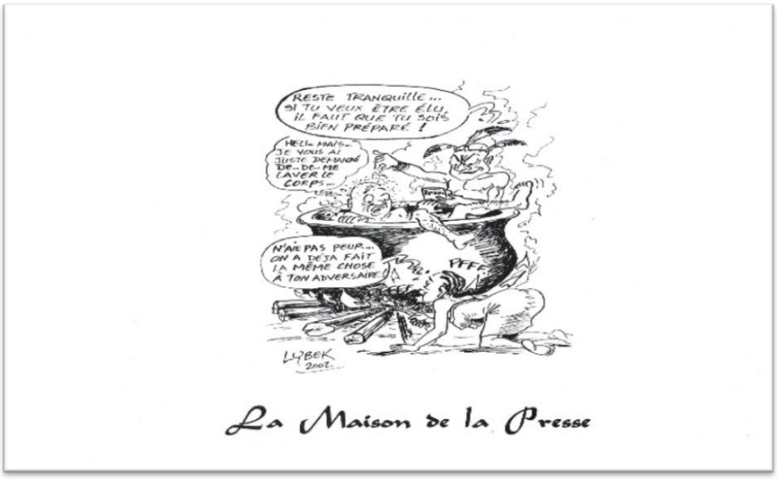

RESTE TRANQUILLE... SI TU VEUX ÊTRE ÉLU, IL FAUT QUE TU SOIS BIEN PRÉPARÉ !
HEU... MAIS... JE VOUS AI JUSTE DEMANDÉ DE... DE... ME LAVER LE CORPS...
N'AIE PAS PEUR... ON A DÉJA FAIT LA MÊME CHOSE À TON ADVERSAIRE!
PFFF
LYBEK 2002
La Maison de la Presse

Figure 7: The back cover of *Gabonitudes Tome I*

Secondly, Gabonese cartoonists play a major role in popular education, particularly in raising awareness about social phenomena or public policies. In this context, the French language in use will be localized although maintaining the same grammatical and orthographic rigor as standard language in formal or common speech. Figure 8 below is a cartoon by Jeff Ikapi raising awareness on the enforcement of a recently signed law against begging.

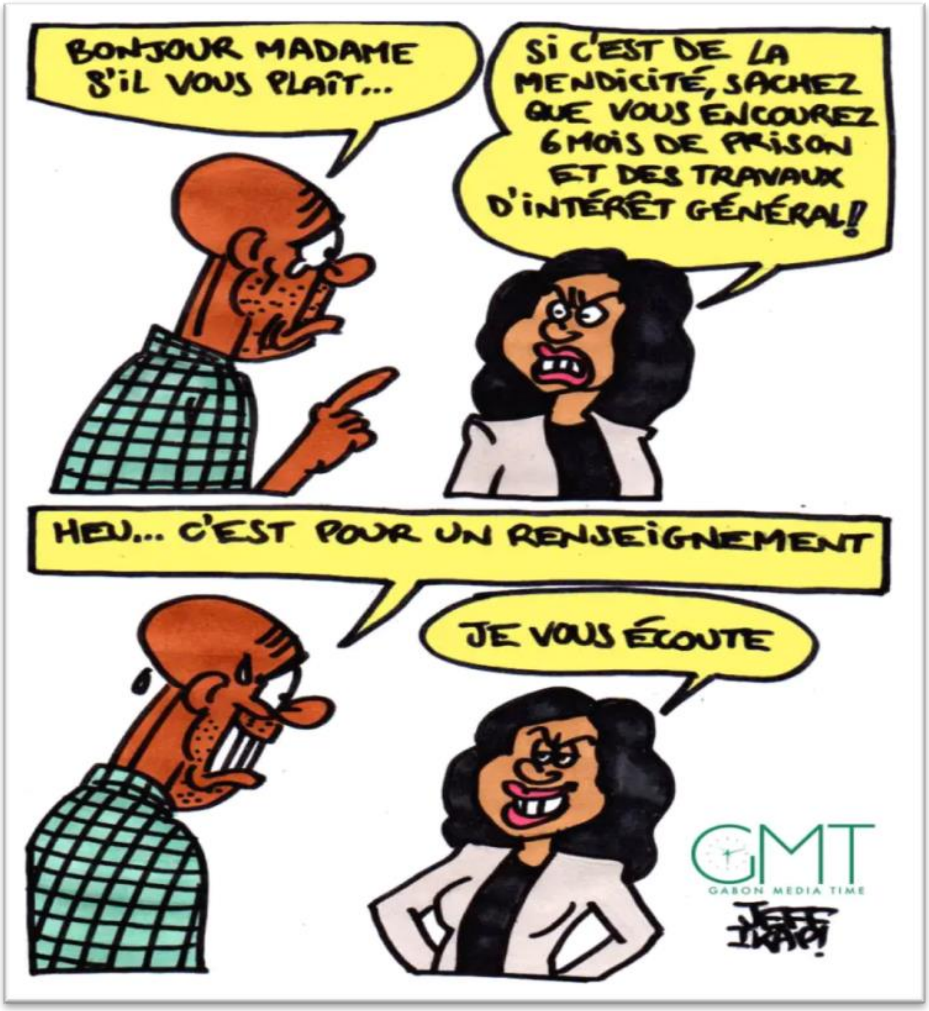

BONJOUR MADAME S'IL VOUS PLAîT...
SI C'EST DE LA MENDICITÉ, SACHEZ QUE VOUS ENCOUREZ 6 MOIS DE PRISON ET DES TRAVAUX D'INTÉRÊT GÉNÉRAL!
HEU... C'EST POUR UN RENSEIGNEMENT
JE VOUS ÉCOUTE
GMT
GABON MEDIA TIME
JEFF IKAPI

Figure 8: Raising awareness of the law against begging (by Jeff Ikapi)[9]

Finally, cartoons, because they are mostly published in the media (online and on paper), are also involved in the popularization of Gabonisms and the circularization of a certain social culture in Gabon.

For instance, Figure 9 below, another work by Jeff Ikapi, while it denounces the endemic unemployment in Gabon, also deplores both the nepotism in job offers and the lack of humanism in friends and acquaintances.

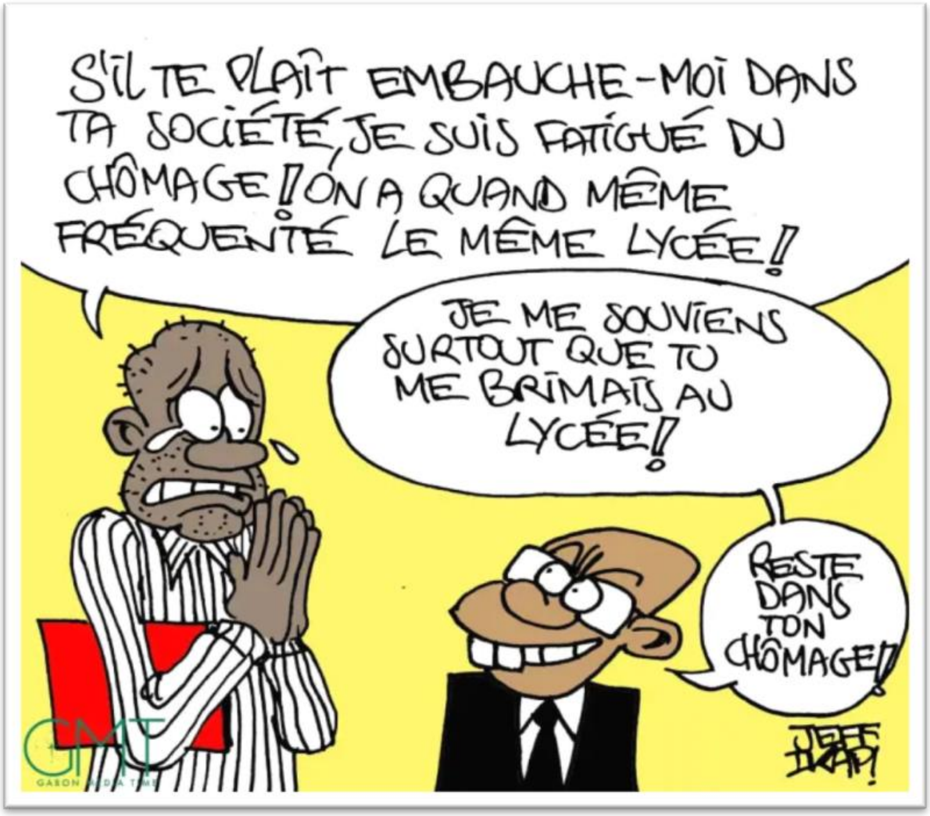

S'IL TE PLAÎT EMBAUCHE-MOI DANS TA SOCIÉTÉ, JE SUIS FATIGUÉ DU CHÔMAGE! ON A QUAND MÊME FRÉQUENTÉ LE MÊME LYCÉE!
JE ME SOUVIENS SURTOUT QUE TU ME BRIMAIS AU LYCÉE!
RESTE DANS TON CHÔMAGE!
JEFF IKAPI
GMT
GABON MEDIA TIME

Figure 9: An aspect of Gabonese social culture (by Jeff Ikapi)[10]

In Figure 10 below, the term "moupohou", a Gabonism, is being propagated. In most languages of Southern Gabon such as Gisir, Yilumbu and Yipunu, the word **moupohou** [mùpóɣù̥] designates a vegetable, consisting of the young leaves of the taro plant (*Colocasia Esculenta*). Like the taro root, the vegetable is very common and popular in Gabonese cuisine. As a Gabonism in French, the term designates the way of making a lot of money easily, and mostly by manipulation or trickery or simply by fraudulent means.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1702 (Article)

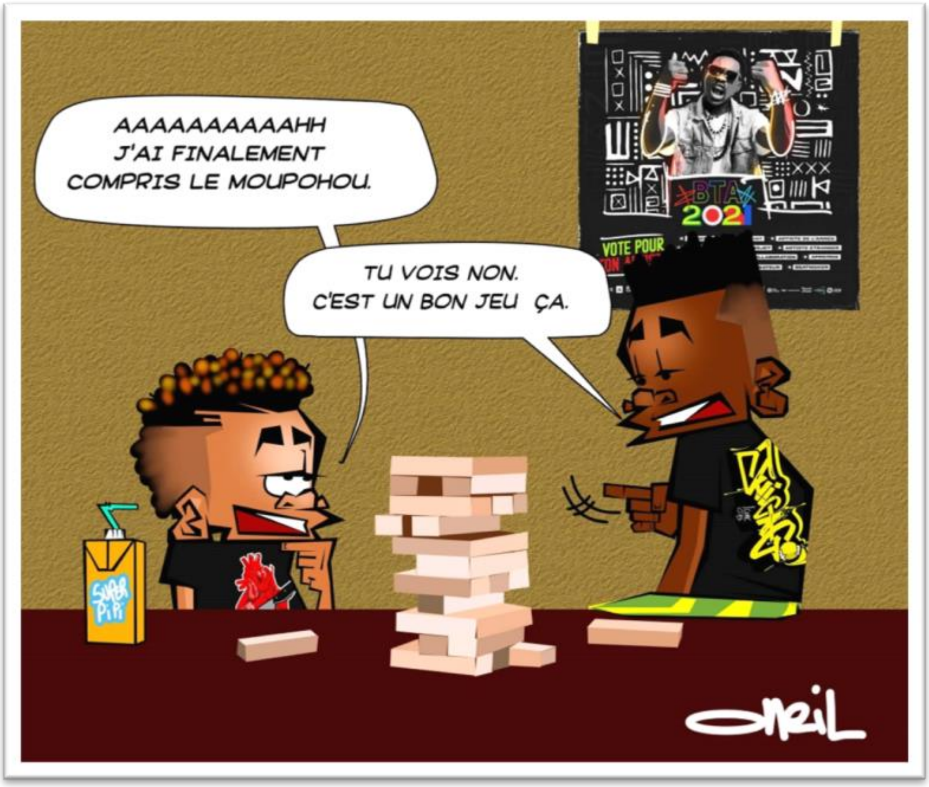


Figure 10: Vehicularization of a Gabonism (by Oneil)[11]

5.2 Primary sources: oral materials

Up to now, we have only considered the collection of written data as part of both the material acquisition phase and the material preparation phase of the planned dictionary project. In many cases, a corpus compiled only from written sources will not be fully representative of the lexical stock of the language (Gouws 2001). That is the reason why we will also collect data (debates, informal face-to-face conversation, etc.) at grassroots level through fieldwork. Government broadcasts in French as well as in the indigenous languages will also provide oral data from *Gabon Télévisions* and *Radio Gabon*, which are owned and operated by the Gabonese government through the Ministry of Communications.

The recordings of the orature as primary sources have been completed for the following provinces: Estuaire (Libreville in particular, see No. 1 in Figure 11), Ogooué-Maritime (Port-Gentil in particular, see No. 8 in Figure 11) and Ngounié (Mouila in particular, see No. 4 in Figure 11).

Number	Province	Capital City
1	Estuaire	Libreville
2	Haut-Ogooué	Franceville
3	Moyen-Ogooué	Lambaréné
4	Ngounié	Mouila
5	Nyanga	Tchibanga
6	Ogooué-Ivindo	Makokou
7	Ogooué-Lolo	Koulamoutou
8	Ogooué-Maritime	Port-Gentil
9	Woleu-Ntem	Oyem

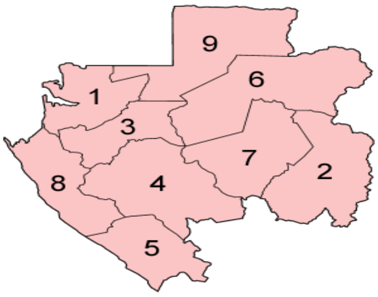


Figure 11: The nine provinces of Gabon[12]

The following provinces have not been covered yet, namely Woleu-Ntem (see No. 9 in Figure 11), Ogooué-Ivindo (see No. 6 in Figure 11), Moyen-Ogooué (see No. 3 in Figure 11), Ogooué-Lolo (see No. 7 in Figure 11), Haut-Ogooué (see No. 2 in Figure 11) and Nyanga (see No. 5 in Figure 11). As far as this point is concerned, an early identification of short, medium and long-term objectives is necessary for future successful data collection.

As part of both the material acquisition phase and the material preparation phase, the prospective editor-in-chief of the planned dictionary project should be responsible for the planning and conducting/overseeing of prospective fieldtrips in order to collect data as well as logistics and managerial aspects to facilitate the material collection phase.

5.3 Secondary and tertiary sources

The secondary sources of the DGFG are all available dictionaries in French as it exists in Gabon. Two kinds of these dictionaries exist in Gabon:

(i) the formal dictionaries produced in France, and

(ii) the dictionaries of localized French produced in the last two decades.

The formal dictionaries produced in France are those that the *Académie Française*[13] currently recognize are the *dictionnaires de la langue française moderne* (dictionaries of modern French language)[14]. The current major dictionaries are produced by French publishing companies such as Larousse, Hachette and Dictionnaires Le Robert. These dictionaries are used in Gabon as premium dictionaries in all domains of life and primarily in the education sector.

The dictionary basis of these dictionaries is indeed the French language as it is spoken in France, and more certainly the Parisian French which is standard language in France. The use in Gabon of these French dictionaries and of the Parisian French as the acrolectal variant of Gabonese French may be at the origin

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1702 (Article)

of the belief that the French language spoken in Gabon is in Africa the closest to Parisian French (Boussougou 2011; Minko 2008).

The dictionaries of localized French are the so-called dictionaries of Gabonese French as successively presented and analyzed in Mavoungou (2013) and Nyangone et al. (2016). The two kinds of dictionaries as well as the tertiary sources (all other linguistic material such as grammars, scientific articles, and books) and the primary sources presented earlier form not only the basis of the projected dictionary but also the initial framework for the corpus of Gabonese French.

6. Conclusion

The discussion above has sought to share some reflections on a few theoretical perspectives for a general dictionary of Gabonese French. The planned dictionary will be a comprehensive one in the sense that it will list words and expressions attested in Gabonese French, namely acrolectal forms (standard French), upper mesolectal forms (common Gabonese French), popular Gabonese French and Gabonese Matitis French. The article has discussed some steps of the dictionary conceptualization plan towards the projected dictionary.

Copyright issues should be clarified before starting with the digitization of all the data that are only available in paper format as well as the use of semi-automatic extraction of terms from all identified online sources. Once data collection and data processing are done, the planned dictionary will be refined and tested in relation to the needs and reference skills of its prospective users. Finally, the planned dictionary as well as the current paper will contribute to laying the foundation for Gabonese French lexicography. Equally, the availability of such a dictionary (published online or in paper format) may be an important step towards the codification of the French language variety of Gabon.

Acknowledgements

The authors wish to express their gratitude to Dr Peter Plüddermann for his comments and suggestions on the pre-final version of this article. The adjudicators' comments and suggestions also contributed to improving this article. All shortcomings, errors and claims remain the sole responsibility of the authors.

Notes

[1] Translated from Mabika Mbokou (2019: 2) "parler français au Gabon, c'est parler un français dont le lexique, le sens des mots et leurs usage est différent de la norme du français standard".

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1702 (Article)

2. GRELACO (*Groupe de Recherche en Langues et Cultures Orales*) is a research unit within the Department of Language Sciences at Omar Bongo University.
3. Source: https://www.union.sonapresse.com/gabon-economie/reduction-du-train-de-vie-de-letat-ce-qui-va-changer-18025. Consulted on 27-06-2018.
4. Source: https://www.union.sonapresse.com/gabon-economie/reduction-du-train-de-vie-de-letat-ce-qui-va-changer-18025. Consulted on 27-06-2018.
5. The digitized versions of most Gabonese newspapers and magazines can be found on the e-kiosk of SOGAPRESSE, which is the only distributor of press publications in Gabon: https://www.e-kiosque-sogapresse.com
6. Source: https://en.wikipedia.org/wiki/Media_of_Gabon. Consulted on 27-06-2018.
7. Source: http://www.gaboneco.com/gabon-pahe-victime-de-ses-caricatures.html. Consulted on 25-07-2019.
8. Source: https://www.bedetheque.com/serie-50587-BD-Gabonitudes.html
9. Source: https://www.gabonmediatime.com/la-mendicite-penalisee
10. Source: https://www.gabonmediatime.com/violences-en-milieu-scolaire-rira-bien-qui-rira-le-dernier
11. Source: https://web.facebook.com/MadLight241/photos/a.1955739911325554/3215597432006456/?type=3&theater
12. Source: https://en.wikipedia.org/wiki/Subdivisions_of_Gabon
13. *Académie Française* or French Academy is France's linguistic watchdog for the French language.
14. Cf. https://www.academie-francaise.fr/les-dictionnaires-du-francais-moderne

References

Dictionaries and encyclopedias

Boucher, K. and S. Lafage. 2000. *Le lexique français du Gabon (entre tradition et modernité). Le Français en Afrique* 14. Special Issue. Nice: Institut de Linguistique Française.

Ditougou, L. 2009. *On est ensemble: 852 mots pour comprendre le français du Gabon.* Libreville: Editions Raponda-Walker.

Dodo-Bounguendza, E. 2008. *Dictionnaire des gabonismes.* Paris: L'Harmattan.

Dodo-Bounguendza, E. 2010. *Diagnostic du français du Gabon. Guide pratique destiné aux journalistes, politiques, administratifs et universitaires.* Libreville: Les Editions Ntsame.

Dodo-Bounguendza, E. 2013. *Dictionnaire du parler toli-bangando. Argot des jeunes gabonais.* Libreville: Les Editions Ntsame.

Mavoungou, P.A., F. Moussounda Ibouanga and J.-A. Pambou. 2014. *Le dico des makaya et des mamadou. Contribution à l'étude du français du Gabon*. Libreville: Editions Odette Maganga.

Mavoungou, P.A., F. Moussounda Ibouanga and J.-A. Pambou. 2015. *Le dico des makaya et des mamadou. Contribution à l'étude du français du Gabon*. Second edition. Libreville: Editions Odette Maganga.

Moussounda Ibouanga, F. 2011. *Français du Gabon: approches sociolinguistiques et lexicographiques (le toli bangando).* Paris: Editions Universitaires Européennes.

Other literature

André, F. 2017. *Pratiques Scripturales et Ecriture SMS. Analyse Linguistique d'un Corpus de Langue Française.* Ph.D. thesis. Paris: Université Paris-Sorbonne.

Andreassen, H.N. 2018. La diérèse en français suisse: Esquisse d'un projet. *Paper presented at the Journées FLORAL-(I)PFC 2018 Conference (Contact de Langues et (Inter)Phonologie de Corpus), Paris, 22–27 November 2018.*

Artigues, M. 1995. *Participation à une étude des particularités lexicales du français parlé au Gabon*. M.A. thesis. Paris: Université Sorbonne Nouvelle.

Atsé N'Cho, J.-B. 2018. Appropriation du français en contexte plurilingue africain: le nouchi dans la dynamique sociolinguistique de la Côte d'Ivoire. *SHS Web of Conferences* 46(4): 13002. Congrès Mondial de Linguistique Française — CMLF 2018. Available at: https://doi.org/10.1051/shsconf/20184613002

Attieh, R., K. Koffi, M. Touré, É. Parr-Labbé, A.H. Pakpour and T.G. Poder. 2022. Validation of the Canadian French Version of the Fear of COVID-19 Scale in the General Population of Quebec. *Brain and Behavior* 12(5): e32550. DOI: 10.1002/brb3.2550.

Benzakour, F. 2010. Le français au Maroc. Enjeux et réalité. *Le Français en Afrique* 25: 33-41.

Bongo Ondimba, O. 1998. *Les Chances du Gabon pour l'An 2000, le Chemin du Futur.* Libreville: Multipress.

Boucher, K. 1997. *Créativité lexicale et identité culturelle du français au Gabon*. Unpublished M.A. thesis. Paris: Université Sorbonne Nouvelle.

Boussougou, S. 2011. *Assessing the Impact of French on the Language Varieties of Gabon*. Ph.D. thesis. Liverpool: Liverpool John Moores University.

Boussougou, S. and K. Menacere. 2015. *The Impact of French on the African Vernacular Languages. For Better or for Worse? Gabon as a Case Study.* Newcastle upon Tyne: Cambridge Scholars Publishing.

Boutin-Dousset, C. 1989. *Matériaux pour un inventaire des particularités lexicales du français au Gabon.* M.A. thesis. Paris: Université Sorbonne Nouvelle.

Diarra, A. 2018. *Le bilinguisme scolaire au Mali: une école qui bidouille entre français et langues nationales.* Paper presented at the Conference on Development and Diversity held in Abidjan, Cote d'Ivoire, 28 November 2018.

Djoum Nkwescheu, A. 2008. Les tendances fédératrices des déviations du français camerounais. De l'identité des processus linguistiques dans les changements diachroniques et géographiques. *Le français en Afrique* 23: 167-198.

Eloundou Eloundou, V. 2019. Le français au Cameroun: constructions socio-identitaires et significativité. Reguigui, A., J. Boissonneault, L. Messaoudi, H. El Amrani and H. Bendahmane (Eds.). 2019. *Langues en Contexte. Languages in Context:* 193-218. Gatineau: Éditions OKAD.

Emejulu, J.D. 2000. Lexicography, an Economic Asset in Multilingual Gabon. *Revue Gabonaise des Sciences du Langage/Gabonese Journal of Language Sciences* 1: 51-69.

Emejulu, J.D. 2001. Lexicographie multilingue et multisectorielle au Gabon: planification, stratégie et enjeux. Emejulu, J.D. (Ed.). 2001. *Eléments de lexicographie gabonaise. Tome I:* 38-57. New York: Jimacs-Hillman.

Emejulu, J.D. 2002. Défis et promesses de la lexicographie intégrale dans les pays en développement. Emejulu, J.D. (Ed.). 2002. *Eléments de lexicographie gabonaise. Tome II:* 366-381. New York: Jimacs-Hillman.

Emejulu, J.D. 2003. Challenges and Promises of a Comprehensive Lexicography in the Developing World: The Case of Gabon. Botha W.F. (Ed.). 2003. *'n Man wat beur. Huldigingsbundel vir Dirk van Schalkwyk:* 195-212. Stellenbosch: Bureau of the WAT.

Equipe DELIC. 2004. Présentation du *Corpus de référence du français parlé*. *Recherches sur le Français Parlé* 18: 11-42.

Fall, O.S. 2021. Histoire et Pratiques du Français au Sénégal. Pour une didactique unifiée des apprentissages. *Revue Etudes Africaines* 3. Available online:
http://webtest.ucad.sn/OJS338/index.php/revueAfricaines/article/view/14

Gouws, R.H. 1999. *A Theoretically Motivated Model for the Lexicographic Processes of the National Lexicography Units.* Research report submitted to the Pan South African Language Board.

Gouws, R.H. 2001. Lexicographic Training: Approaches and Topics. Emejulu, J.D. (Ed.). 2001: *Eléments de lexicographie gabonaise. Tome I:* 58-94. New York: Jimacs-Hillman.

Gouws, R.H. 2003. Towards the Formulation of a Theoretically Motivated Model for the National Lexicography Units in South Africa. Hartmann, R.R.K. (Ed.). 2003. *Lexicography: Critical Concepts*: 218-246. London/New York: Routledge.

Gouws, R.H. and D.J. Prinsloo. 2005. *Principles and Practice of South African Lexicography*. Stellenbosch: SUN PReSS.

Hambye, P. and A.C. Simon. 2012. The Variation of Pronunciation in Belgian French: from Segmental Phonology to Prosody. Gess, R., C. Lyche and T. Meisenburg (Eds.). 2012. *Phonological Variation in French: Illustrations from Three Continents:* 129-149. Amsterdam: John Benjamins.

Italia, M. 2006. Le morphème *là* dans les variétés mésolectales et basilectales en français du Gabon. *Le Français en Afrique* 21: 281-290.

Kouadio N'Guessan, J. 2008. Le français en Côte d'Ivoire: de l'imposition à l'appropriation décomplexée d'une langue exogène. *Documents pour l'Histoire du Français Langue Étrangère ou Seconde* 40–41: 179-197.

Kumalo, M.B. 1999. The National Lexicography Units — Existing and Prospective. *Lexikos* 9: 211-216.

Mabika Mbokou, L. 2008. Le français langue maternelle! *CENAREST Infos* 4:4.

Mabika Mbokou, L. 2012. A Survey of Bilingualism in Multilingual Gabon. Ndinga-Koumba-Binza, H.S. and S.E. Bosch (Eds.). 2012. *Language Science and Language Technology in Africa. Festschrift for Justus C. Roux:* 163-175. Stellenbosch: SUN PReSS.

Mabika Mbokou, L. 2019. *Le français du Gabon: Mythe ou Réalité?* Unpublished manuscript. Available online:
https://www.researchgate.net/publication/306058155_Le_Francais_du_Gabon_mythe_ou_realite. Consulted on 26 July 2019.

Madiba, M. and D. Nkomo. 2010. The *Tshivenḓa-English Ṱahalusamaipfi/Dictionary* as a Product of South African Lexicographic Processes. *Lexikos* 20: 307-325.

Martineau, F. 2005. Perspectives sur le changement linguistique: aux sources du français canadien. *Canadian Journal of Linguistics/Revue Canadienne de Linguistique* 50(1–4): 173-213.

Martineau, F. 2007. Variation in Canadian French Usage from the 18th to the 19th Century. *Multilingua: Journal of Cross-Cultural and Interlanguage Communication* 26(2–3): 203-227.

Massinga Kombila, M. 2013. *Le Français au Gabon: Émergence d'une Norme Endogène. Le Cas de la Presse Écrite*. Ph.D. thesis. Bordeaux: Université Michel Montaigne (Bordeaux III).

Mavoungou, P.A. 2002. Vers un dictionnaire du français du Gabon. Emejulu, J.D. (Ed.). 2002. *Eléments de Lexicographie Gabonaise*. *Tome II:* 230-262. New York: Jimacs-Hillman.

Mavoungou, P.A. 2010. *Lexicographie et Confection des Dictionnaires au Gabon.* Stellenbosch: SUN PReSS.

Mavoungou, P.A. 2011. Regard sur les onomastismes dans le français de Libreville et leur traitement lexicographique. *Mbaandʒa. Revue d'Etude et d'Analyse Francophones* 1: 21-50.

Mavoungou, P.A. 2013. Gabonese French Dictionaries: Survey and Perspectives. *Lexikos* 23: 255-272.

Mavoungou, P.A., T. Afane Otsaga, G.-R. Mihindou and B. Nyangone Assam. 2002. *Vers un dictionnaire du français du Gabon. Premier dictionnaire parallèle contenant une centaine d'articles traités*. Unpublished document. Stellenbosch: University of Stellenbosch.

McArthur, T. 2006. The Power of Words: Pressure, Prejudice and Politics in our Vocabularies and Dictionaries. Bolton, K. and B.B. Kachru (Eds.). 2006. *World Englishes. Critical Concepts in Linguistics. Vol. 3:* 335-347. London/New York: Routledge.

Mengara, D.M. 2000. French: An African Language, Finally! Carstens, V. and F. Parkinson (Eds). 2000. *Advances in African Linguistics:* 281-298. Trenton, NJ/Asmara: Africa World Press.

Mindze M'Eyeghe, J. 2001. *Approches de quelques particularités lexicales du français parlé au Gabon*. M.A. thesis. Libreville: Omar Bongo University.

Minko, D. 2008. Le marquage identitaire dans le français gabonais. *Synergies Monde* 5: 159-164.

Mitchell, R. 2004. Les perceptions du français gabonais et la distribution des langues au Gabon. *Le Français en Afrique* 19: 177-188.

Mongwe, M.J. 2006. *The Role of the South African National Lexicography Units in the Planning and Compilation of Multifunctional Bilingual Dictionaries.* Unpublished M.Phil. thesis. Stellenbosch: Stellenbosch University.

Mouélé, M. 2011. Les racines bantu du français gabonais. *Mbaandʒa. Revue d'Etude et d'Analyse Francophones* 1: 87-111.

Mougeon, R. and E. Beniak. 1989. *Le Français Canadien Parlé hors Québec: Aperçu Sociolinguistique*. Québec: Les Presses de l'Université Laval.

Mouloungui Nguimbyt, F.V. 2002. *De la variation dialectale en français du Gabon*. Unpublished M.A. thesis. Libreville: Omar Bongo University.

Moussirou Mouyama, A. 1984. *La Langue Française au Gabon: Contribution Sociolinguistique.* Unpublished doctoral thesis. Paris : Université René Descartes (Paris V).

N'Diaye Corréard, G. 2008. Défense et illustration d'un français périphérique: Les mots du patrimoine: le Sénégal. Bavoux, C. (Ed.). 2008. *Le français des Dictionnaires. L'Autre Versant de la Lexicographie Française:* 205-218. Louvain-la-Neuve: De Boeck Supérieur.

Ndinga-Koumba-Binza, H.S. 2005. Considering a Lexicographic Plan for Gabon within the Gabonese Language Landscape. *Lexikos* 15: 132-150.

Ndinga-Koumba-Binza, H.S. 2006. *Lexique Pove–Français/Français–Pove*, Mickala Manfoumbi: Seconde Note de Lecture. *Lexikos* 16: 293-308.

Ndinga-Koumba-Binza, H.S. 2007. Gabonese Language Landscape: Survey and Perspectives. *South African Journal of African Languages* 27(3): 97-116.

Ndinga-Koumba-Binza, H.S. 2011. From Foreign to National: A Review of the Status of the French Language in Gabon. *Literator* 32(2): 135-150.

Ndinga-Koumba-Binza, H.S. 2013. Identité et nationalisation des langues au Gabon. *Mbaandʒa: Revue d'Etude et d'Analyse Francophones* 2: 147-163.

Nel, M.-L. and K. Ferreira-Meyers. 2020. Decolonizing the 'French as a Foreign Language' University Curriculum through Literature, Interpretation from a South(ern) African Perspective. *HyperCultura* 9: 1-14.

Nifaoui, A. 2021. Le statut du français au Maroc face à l'hégémonie de l'anglais: attitudes des apprenants envers le français et l'anglais. *Journal of Applied Language and Culture Studies* 4: 121-142.

Nsa Ndo, E.K. 2010. *Les locutions figurées dans le français des jeunes librevillois de 15 à 30 ans: approche linguistique et sociolinguistique.* M.A. thesis. Libreville: Omar Bongo University.

Nsafou, V. 2010. *Les collocations dans le français des jeunes librevillois de 15 à 30 ans: approche linguistique et lexicologique.* M.A. thesis. Libreville: Omar Bongo University.

Nyangone Assam, B., H.S. Ndinga-Koumba-Binza and V. Ompoussa. 2016. What French for Gabonese French Lexicography? *Lexikos* 26: 162-192.

Ondo Mébiame, P. and G.M. Ekwa Ebanéga. 2011. Regard critique sur *On est ensemble: 852 mots pour comprendre le français du Gabon*. *Lexikos* 21: 337-358.

Pambou, J.-A. 1998. Le Français au Gabon: une langue à multiples statuts. *iBoogha* 2: 127-149.

Pedraza, A.P. and L.-A. Cougnon. 2021. Exploring the Expression of Difficulty in the Belgian French Twitter Corpus of Climate Change. Hoelbeek, T. and L. Rosseel (Eds). 2021. *Linguists' Day of the Linguistic Society of Belgian, Vrije Universiteit Brussel, 22 October 2021. Book of Abstracts:* 47-48. Brussels: Vrije Universiteit Brussel.

Pigeon, J. 2021. *Guide Pratique de la Localisation Vidéoludique en Français Canadien, Précédé de l'État des Lieux*. M.A. thesis. Gatineau: Université du Québec en Outaouais.

Plahar, B. 2017. *Le français en Côte d'Ivoire: Une Analyse Linguistique de Six Animations Ivoiriennes en Français Normé Ivoirien, en Français Populaire Ivoirien et en Nouchi.* M.A. thesis. Ottawa: Carleton University.

Poder, T.G., J.R. Guertin, M. Touré, G. Pratte, C. Gauvin, D. Feeny, W. Furlong and C. Camden. 2021. Canadian French Translation and Linguistic Validation of the Health-related Quality of Life Utility Measure for Pre-school Children. *Expert Review of Pharmacoeconomics & Outcomes Research* 21(6): 1195-1201.

Poliquin, G.C. 2006. *Canadian French Vowel Harmony*. Ph.D. thesis. Cambridge, MA: Harvard University.

Racine, I. and H.N. Andreassen. 2012. A Phonological Study of a Swiss French Variety: Data from the Canton of Neuchâtel. Gess, R., C. Lyche and T. Meisenburg (Eds.). 2012. *Phonological Variation in French: Illustrations from Three Continents:* 173-207. Amsterdam: John Benjamins.

Sertling Miller, J. 2007. *Swiss French Prosody: Intonation, Rate, and Speaking Style in the Vaud Canton.* Unpublished Ph.D. thesis. Urbana-Champaign: University of Illinois at Urbana-Champaign.

Siepmann, D., C. Bürgel and S. Diwersy. 2016. Le *Corpus de référence du français contemporain* (CRFC), un corpus massif du français largement diversifié par genres. *SHS Web of Conferences* 27 11002. DOI: 10.1051/shsconf/20162711002.

Skattum, I. 2010. Le français parlé du Mali: une variété régionale? Abecassis, M. and G. Ledegen (Eds). 2010. *Les Voix des Français, en Parlant, en Écrivant. Vol. 2:* 433-448. Bern: Peter Lang.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1702 (Article)

Svensén, B. 2009. *A Handbook of Lexicography: The Theory and Practice of Dictionary Making.* Cambridge, UK: Cambridge University Press.

Walker, D.C. 1984. *The Pronunciation of Canadian French*. Ottawa: University of Ottawa Press.

Wiegand, Herbert Ernst. 1998. *Wörterbuchforschung. Untersuchungen zur Wörterbuchbenutzung, zur Theorie, Geschichte, Kritik und Automatisierung der Lexikographie.* Volume 1. Berlin/New York: Walter de Gruyter.

Zabus, C. 2007. *The African Palimpsest. Indigenization of Language in the West African Europhone Novel.* Amsterdam/New York: Rodopi.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1703 (Article)

A Lexicographical Perspective to Intentional and Incidental Learning: Approaching an Old Question from a New Angle

Sven Tarp, *Centre for Lexicographical Studies, Guangdong University of Foreign Studies, China; Department of Afrikaans and Dutch, Stellenbosch University, South Africa; International Centre for Lexicography, Valladolid University, Spain; Centre of Excellence in Language Technology, Ordbogen A/S, Denmark, and Centre for Lexicography, Aarhus University, Denmark* (*st@cc.au.dk*)

Abstract: Second-language learning is a complex process that combines text reception (reading, listening) and text production (writing, talking). Applied linguistics usually distinguishes between intentional and incidental learning. The academic literature contains various definitions of these concepts, especially in connection with reading. The paper explores L2 learning from a lexicographical perspective and redefines the two terms based on parameters like flow, focus, and interruption. It then focuses on digital dictionaries integrated into e-readers, learning apps, and writing assistants, and argues that this integration, so far, has not been particularly successful due to a number of negative factors. As an alternative, the paper provides examples of how lexicographical data could be filtered and presented in pop-up windows to serve both incidental and intentional learning. The former requires instantaneous, contextualized, and discreet assistance with an absolute minimum of lexicographical data, whereas the latter presupposes easy access to relevant additional data. Finally, the paper discusses the techniques and technologies required to guarantee this approach.

Keywords: INCIDENTAL LEARNING, INTENTIONAL LEARNING, INTEGRATED DICTIONARIES, E-READERS, E-READING TOOLS, LEARNING APPS, WRITING ASSISTANTS, INTUITIVE USE, CONTEXT-AWARENESS, LEXICOGRAPHICAL CONTEXTUALIZATION

Opsomming: 'n Leksikografiese perspektief op doelbewuste en toevallige leer: 'n Ou vraagstuk word vanuit 'n nuwe invalshoek benader. Die aanleer van 'n tweede taal is 'n komplekse proses waarin teksresepsie (lees, luister) en teksproduksie (skryf, praat) gekombineer word. Die Toegepaste linguistiek onderskei gewoonlik tussen doelbewuste en toevallige leer. Die akademiese literatuur bevat verskeie definisies van hierdie konsepte, veral met betrekking tot lees. In hierdie artikel word L2-leer vanuit 'n leksikografiese perspektief verken, en die twee terme word, gebaseer op parameters soos vloei, fokus, en onderbreking, geherdefinieer. Dan word daar gefokus op digitale woordeboeke wat in e-lesers, aanleerdertoepassings en skryfhulpmiddels geïntegreer is, en daar word geargumenteer dat hierdie integrasie tot dusver weens 'n aantal negatiewe faktore nie besonder suksesvol was nie. As alternatief verskaf hierdie artikel voorbeelde van hoe leksikografiese data gefilter en in opwipvensters tot voordeel van beide toevallige en doelbewuste leer aangebied kan word. Eersgenoemde vereis onmiddellike, gekontekstualiseerde, en diskrete ondersteuning met 'n absolute minimum leksikografiese data, terwyl laasgenoemde maklike toegang tot relevante addisionele data voorveronderstel. Laastens word tegnieke en tegnologieë wat vereis word om die sukses van hierdie benadering te waarborg, bespreek.

Sleutelwoorde: TOEVALLIGE LEER, DOELBEWUSTE LEER, GEÏNTEGREERDE WOORDEBOEKE, E-LESERS, E-LEESHULPMIDDELS, AANLEERDERSTOEPASSINGS, SKRYFHULPMIDDELS, INTUÏTIEWE GEBRUIK, KONTEKSBEWUSTHEID, LEKSIKOGRAFIESE KONTEKSTUALISERING

1. Introduction

Second-language (L2) learning is a complex process that combines text reception (reading, listening) and text production (writing, talking). In the linguistic and psycho-linguistic tradition, the language-learning process is usually separated into intentional learning and incidental learning; see Krashen (1989), Nagy et al. (1985), Bereiter and Scardamalia (1989), Hulstijn (1989, 2013), Shu, Anderson and Zhang (1995), Laufer and Hulstijn (2001), Brown, Waring and Donkaewbua (2008), and Leow and Zamora (2017), among many others.

The academic literature provides a large number of definitions of these terms, especially in connection with reading. Sometimes, the two terms are treated as synonymous to explicit and implicit learning, respectively, but this approach has been questioned by other authors like Hulstijn (1989: 2633). In any case, the distinctive criterion in the various definitions seems to be the existence or lack of *intent to learn*, frequently in combination with the learner's awareness and consciousness of the process. Hulstijn (1989: 2632) himself defines intentional learning as "a deliberate attempt to commit factual information to memory". Bereiter and Scardamalia (1989: 363) relate it to the "cognitive processes that have learning as a goal rather than an incidental outcome". In this connection, Leow and Zamora (2017: 33) comment that intentional learning "has always been assumed to represent the type of learning, of a more explicit nature, that underscores a formal instructional classroom setting". According to the two authors, the definition of this concept is "relatively stable … albeit with some nuances" whereas there is "quite a range of perceptions" of what incidental learning entails. The different perceptions are "typically reflected in the methodology employed to address its role in the L2 learning process".

Without going too deep into this discussion, a methodological problem in existing research seems to be the lack of terminological distinction between *language knowledge* and *language skills*; see Tarp (2008: 131-136). After reading a text, informants are typically asked what they *know* about a word in terms of meaning, gender, morphosyntax, etc. However, and as Lessing (1747: 8-9) helped us to understand almost three hundred years ago, the purpose of second-language learning is not to obtain some kind of (learned) knowledge of this language, but to develop language skills, that is, "the ability to communicate in the language concerned: to read, write, listen to and speak this language" (Tarp 2008: 132). Learners should not only know words, they should also be able to use them in real life.

From this perspective, another methodological challenge is that many (most?) studies of incidental learning have been conducted in the classroom, or the school context in general, in connection with the informants' reading of written texts. This, of course, provides a more controlled environment to extract reliable empirical data and reach science-based conclusions. But it also excludes other types of social contexts where learners engage in written and oral communication and incidentally pick up words, meanings, and grammatical structures. These contexts also play a relevant role, and sometimes even a crucial one, in second-language learning, especially when the process occurs inside the geographical area where the concerned language is the dominant one. Thus, if the overall learning process has to be grasped in its totality, it seems equally relevant to relate incidental learning to the text production process, especially writing, and to oral communication in general. According to the newest research, modern human beings have existed for about 300,000 years, whereas written language has a much shorter history of 5,000 years for the privileged few and less than two hundred years for the vast majority. This suggests that spoken language has been the primary means of communication for most of their existence. To the extent our ancestors have learned a second or third language, this must have happened incidentally without written texts and formal instruction. Hence, a short excursion into oral communication may be relevant if we want to achieve a broader perspective of the phenomenon of incidental learning.

In the next section, this idea will be used as a background to explore how lexicography can take advantage of current technologies and add a new dimension to the discussion of intentional and incidental L2 learning in relation to digital devices. Section 3 will then look at dictionaries integrated into e-reading tools (e-readers and similar devices) and explore to which degree digital technologies allow us to offer the "ideal solution". Section 4 will follow up with a short discussion of dictionaries that are integrated into learning apps and provide assistance to the reading of L2 texts. The discussion will show how the application of available technologies and techniques already makes allowance for the "ideal solution". Section 5 will deal with multi-word units of meaning and discuss how this challenge can be treated lexicographically in the tools discussed in the two previous paragraphs and, thus, contribute to the two types of learning. Section 6 will move from reading to writing and briefly discuss how digital writing assistants also can contribute to incidental and intentional learning if lexicographical handicraft is combined with cutting-edge technology. Finally, Section 7 will contain the conclusions and sum up how and to which degree lexicography can contribute to incidental and intentional learning.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1703 (Article)

2. A lexicographical perspective

In recent years, lexicographers have increasingly emphasised the close relationship between lexicography and information science, and some of them have even categorised the former "as part of" the latter (Wiegand 2013: 14). They have good reasons to do so. The core purpose of all lexicographical products is to assist users with *information* that can meet their needs in different types of situations, among them the ones the Function Theory classifies as communicative situations (e.g. text production and text reception); see Fuertes-Olivera and Tarp (2014: 52).

In the following, we will explore how the provision of information may enhance the learning process in terms of incidental and intentional learning. We will start with incidental learning and illustrate it by means of three examples of oral communication that are well-known to most L2 learners, either inside or outside the classroom, either inside or outside the geographical area where L2 is spoken as a native language.

In the first example (Figure 1), an L2 learner is listening to a native speaker who is describing an experience from the previous day. The learner does not understand one of the words used by the native speaker and asks him to explain its meaning.

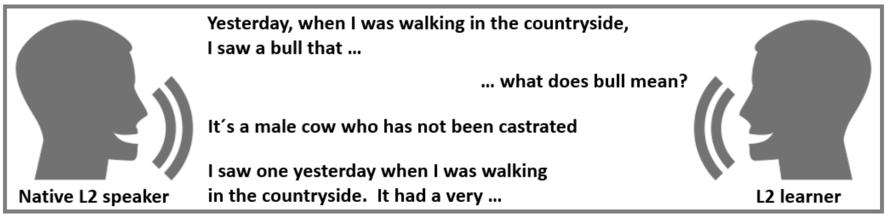

Yesterday, when I was walking in the countryside, I saw a bull that …
… what does bull mean?
It´s a male cow who has not been castrated
I saw one yesterday when I was walking in the countryside. It had a very …
Native L2 speaker
L2 learner

Figure 1: Oral communication between native L2 speaker and L2 learner

In the second example (Figure 2), the roles have shifted. It is now the learner who is describing an experience he had the previous day. Suddenly, he lacks a central word and uses a paraphrase to ask the native speaker to provide the right word.

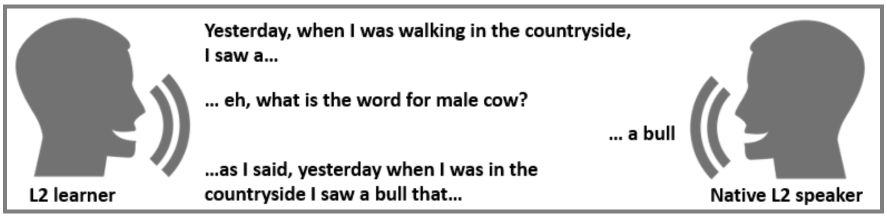

Yesterday, when I was walking in the countryside, I saw a…
… eh, what is the word for male cow?
… a bull
…as I said, yesterday when I was in the countryside I saw a bull that…
L2 learner
Native L2 speaker

Figure 2: Oral communication between L2 learner and native L2 speaker

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1703 (Article)

The third example (Figure 3) shows the learner explaining the same experience, but this time to an L1-speaking teacher of L2. When he does not know or remember a word to express what he wants to say, he therefore addresses the teacher in his mother tongue. The conversation then becomes bilingual and resembles the consultation of L1–L2 dictionaries in connection with L2-text production.

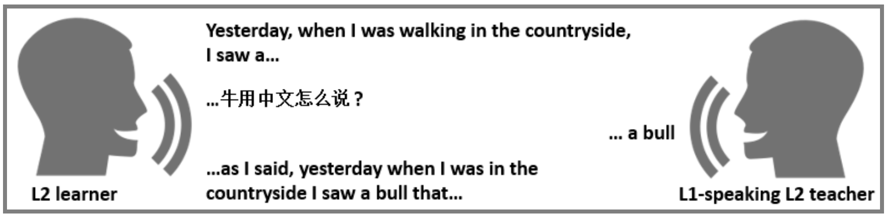

Yesterday, when I was walking in the countryside,
I saw a...
...牛用中文怎么说？
... a bull
...as I said, yesterday when I was in the
countryside I saw a bull that...
L2 learner
L1-speaking L2 teacher

Figure 3: Oral communication between L2 learner and L1-speaking L2 teacher

What do the three figures show? First, they confirm the importance of relating text production and text reception in the language-learning process. Secondly, and apart from the trivial story reproduced, they represent three different types of situations inherent to the topic under discussion. In all of them, the oral conversation between two persons is shortly interrupted. One of the persons does not understand an L2 word in the specific context or does not know which L2 word to use to express his ideas. The other person is then used as a human information resource (human dictionary) that is consulted to get the appropriate information. The short interruption in the conversation fits naturally into the mainline of communication without any of the two persons losing focus on the topic discussed.

The three situations depicted in Figures 1–3 are examples where incidental learning may happen enhanced by the immediate provision of a small piece of information when a communication problem occurs. Although the conversation takes place in the overall framework of the L2-learning process, the learner's *primary intention* in at least two of the three situations is not to learn the second language but to enjoy a normal social conversation.

Inspired by the above examples of oral communication, the concepts of incidental and intentional learning can be adapted to lexicographical consultations performed in connection with the production and reception of written texts. Incidental learning is here related to the situation where learners experience an information need, look up in dictionaries, and get an immediate response that allows them to maintain the reading or writing flow without losing focus on the text and its content and, in this way, pick up new words and meanings. By contrast, intentional learning only starts when the learners interrupt the reading or writing process in order to dedicate time to a deeper study of words, senses, or grammatical structures appearing in the text.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1703 (Article)

If the various types of informal chatting are excluded, the writer and the reader are usually separated in time and only interact directly with the dictionary when engaging in communication by means of written texts (see Figure 4). Hence, to achieve incidental learning from the perspective of lexicography, the challenge is to make the consultation process as easy and smooth as possible and reduce the consultation time to an absolute minimum.

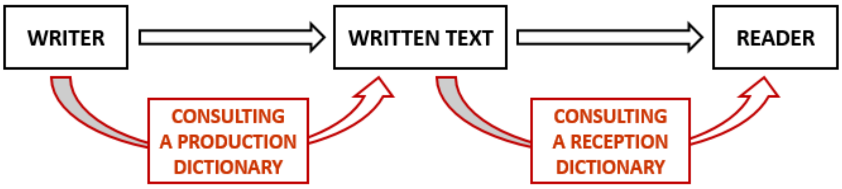

WRITER
WRITTEN TEXT
READER
CONSULTING A PRODUCTION DICTIONARY
CONSULTING A RECEPTION DICTIONARY

Figure 4: Communication by means of written texts

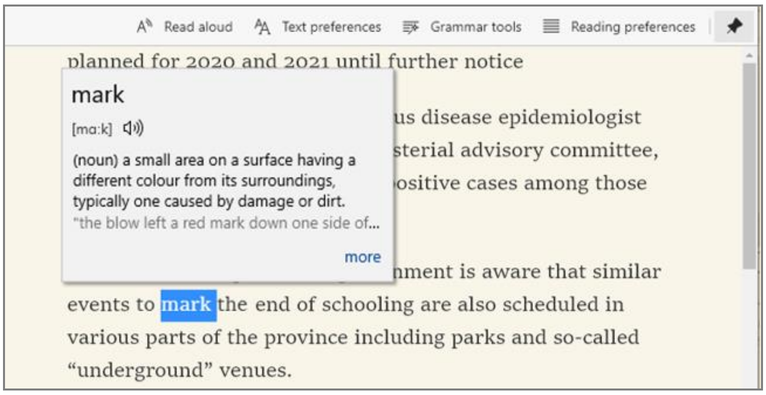
What happens when learners consult traditional dictionaries, whether printed or app-based ones? First and foremost, they will have to leave the text they are reading or writing and consult an external lexicographical resource. Here, they will frequently get access to a complete article with a large amount of lexicographical data, most of which are totally irrelevant in the concrete context, i.e. information overload. This kind of consultation takes time, disturbs the workflow, and hampers incidental learning. The learners' focus moves from the communication to the consultation.

Hence, the ideal solution seems to be dictionaries that have been integrated into the digital devices learners increasingly use when they read and write. Such dictionaries allow the learners to perform lexicographical consultations without leaving the text they are working with and can be found in e-readers, learning apps, and writing assistants, among others. However, the integration of dictionaries into these devices has not, so far, been particularly successful due to conservative thinking, inadequate adaptation to different types of user needs, imperfect lexicographical databases, poor design of user interfaces, and insufficient application of the available technology. In the following, we will look at both existing integrated dictionaries and those to come.

3. Dictionaries integrated into e-readers and similar devices

We will start with the dictionaries integrated into e-reading tools such as e-readers and similar devices used to read digital texts (e.g. laptops, tablets, and smartphones). These dictionaries are probably the most well-known to users and, at the same time, those that present the biggest challenges in terms of smart technology. They have already been discussed and criticised by many researchers, among them various South African lexicographers, incl. Danie Prinsloo, to whom this article is dedicated; see Bothma and Prinsloo (2013), and Bothma and Gouws (2020).

Compared to traditional dictionaries, the ones integrated into other devices have the advantage that they can be activated by simply touching or clicking on the screen. This technique shortens the initial consultation time and implies that users do not have to leave the book or article they are reading. So far, so good! The problems start when a dictionary article is uploaded to the screen. As an illustration, we have chosen a newspaper article about coronavirus published in Times Live (Singh 2020) and displayed on a laptop. At the end of the article, we have clicked on the word *mark* and activated the integrated dictionary, which is "powered by Oxford Dictionaries". The result is the visualisation of a small excerpt of lexicographical data from the corresponding article (see Figure 5).



planned for 2020 and 2021 until further notice

mark

[mɑ:k]

(noun) a small area on a surface having a different colour from its surroundings, typically one caused by damage or dirt.
"the blow left a red mark down one side of...

more

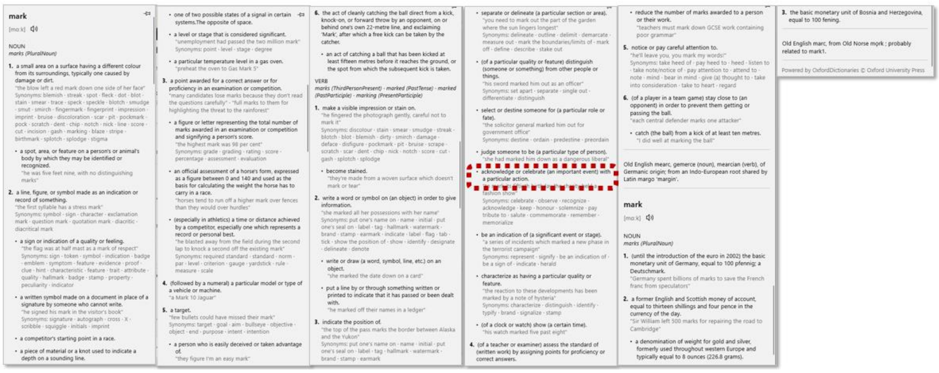
us disease epidemiologist
sterial advisory committee,
ositive cases among those

nment is aware that similar
events to mark the end of schooling are also scheduled in various parts of the province including parks and so-called “underground” venues.

Figure 5: Pop-up window activated on laptop while reading on Internet

It is undoubtedly a good idea that the default pop-up window shown in Figure 5 only furnishes a few lexicographical data. It makes it easier to overview. However, the word *mark* clicked on in the text is a verb, whereas the visualised lemma is a noun. The reader therefore has to make a second click, this time on the signifier "more", to search for the relevant data. The result is a very long article where the user has to scroll down several times. Figure 6 gives an overview of this article after being remodeled. It contains two nouns (based on etymological criteria) and a verb, all of them with two or more senses. The dotted circle highlights the only data relevant in the concrete context. The remaining data are completely superfluous. It will probably take some time to find the required data. As a consequence, the reader's focus will move from the text to the consultation, the reading flow will be interrupted, and incidental learning as defined above becomes impossible. Something has to be done!

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1703 (Article)



mark

[mɑːk]

NOUN
marks (PluralNoun)

1. a small area on a surface having a different colour from its surroundings, typically one caused by damage or dirt.
"the blow left a red mark down one side of her face"
Synonyms: blemish · streak · spot · fleck · dot · blot · stain · smear · trace · speck · speckle · blotch · smudge · smut · smirch · fingermark · fingerprint · impression · imprint · bruise · discoloration · scar · pit · pockmark · pock · scratch · dent · chip · notch · nick · line · score · cut · incision · gash · marking · blaze · stripe · birthmark · splotch · splodge · stigma
   - a spot, area, or feature on a person's or animal's body by which they may be identified or recognized.
"he was five feet nine, with no distinguishing marks"

2. a line, figure, or symbol made as an indication or record of something.
"the first syllable has a stress mark"
Synonyms: symbol · sign · character · exclamation mark · question mark · quotation mark · diacritic · diacritical mark
   - a sign or indication of a quality or feeling.
"the flag was at half mast as a mark of respect"
Synonyms: sign · token · symbol · indication · badge · emblem · symptom · feature · evidence · proof · clue · hint · characteristic · feature · trait · attribute · quality · hallmark · badge · stamp · property · peculiarity · indicator
   - a written symbol made on a document in place of a signature by someone who cannot write.
"he signed his mark in the visitor's book"
Synonyms: signature · autograph · cross · X · scribble · squiggle · initials · imprint
   - a competitor's starting point in a race.
   - a piece of material or a knot used to indicate a depth on a sounding line.
   - one of two possible states of a signal in certain systems.The opposite of space.
   - a level or stage that is considered significant.
"unemployment had passed the two million mark"
Synonyms: point · level · stage · degree
   - a particular temperature level in a gas oven.
"preheat the oven to Gas Mark 5"

3. a point awarded for a correct answer or for proficiency in an examination or competition.
"many candidates lose marks because they don't read the questions carefully" · "full marks to them for highlighting the threat to the rainforest"
   - a figure or letter representing the total number of marks awarded in an examination or competition and signifying a person's score.
"the highest mark was 98 per cent"
Synonyms: grade · grading · rating · score · percentage · assessment · evaluation
   - an official assessment of a horse's form, expressed as a figure between 0 and 140 and used as the basis for calculating the weight the horse has to carry in a race.
"horses tend to run off a higher mark over fences than they would over hurdles"
   - (especially in athletics) a time or distance achieved by a competitor, especially one which represents a record or personal best.
"he blasted away from the field during the second lap to knock a second off the existing mark"
Synonyms: required standard · standard · norm · par · level · criterion · gauge · yardstick · rule · measure · scale

4. (followed by a numeral) a particular model or type of a vehicle or machine.
"a Mark 10 Jaguar"

5. a target.
"few bullets could have missed their mark"
Synonyms: target · goal · aim · bullseye · objective · object · end · purpose · intent · intention
   - a person who is easily deceived or taken advantage of.
"they figure I'm an easy mark"

6. the act of cleanly catching the ball direct from a kick, knock-on, or forward throw by an opponent, on or behind one's own 22-metre line, and exclaiming 'Mark', after which a free kick can be taken by the catcher.
   - an act of catching a ball that has been kicked at least fifteen metres before it reaches the ground, or the spot from which the subsequent kick is taken.

VERB
marks (ThirdPersonPresent) · marked (PastTense) · marked (PastParticiple) · marking (PresentParticiple)

1. make a visible impression or stain on.
"he fingered the photograph gently, careful not to mark it"
Synonyms: discolour · stain · smear · smudge · streak · blotch · blot · blemish · dirty · smirch · damage · deface · disfigure · pockmark · pit · bruise · scrape · scratch · scar · dent · chip · nick · notch · score · cut · gash · splotch · splodge
   - become stained.
"they're made from a woven surface which doesn't mark or tear"

2. write a word or symbol on (an object) in order to give information.
"she marked all her possessions with her name"
Synonyms: put one's name on · name · initial · put one's seal on · label · tag · hallmark · watermark · brand · stamp · earmark · indicate · label · flag · tab · tick · show the position of · show · identify · designate · delineate · denote
   - write or draw (a word, symbol, line, etc.) on an object.
"she marked the date down on a card"
   - put a line by or through something written or printed to indicate that it has passed or been dealt with.
"he marked off their names in a ledger"

3. indicate the position of.
"the top of the pass marks the border between Alaska and the Yukon"
Synonyms: put one's name on · name · initial · put one's seal on · label · tag · hallmark · watermark · brand · stamp · earmark
   - separate or delineate (a particular section or area).
"you need to mark out the part of the garden where the sun lingers longest"
Synonyms: delineate · outline · delimit · demarcate · measure out · mark the boundaries/limits of · mark off · define · describe · stake out
   - (of a particular quality or feature) distinguish (someone or something) from other people or things.
"his sword marked him out as an officer"
Synonyms: set apart · separate · single out · differentiate · distinguish
   - select or destine someone for (a particular role or fate).
"the solicitor general marked him out for government office"
Synonyms: destine · ordain · predestine · preordain
   - judge someone to be (a particular type of person).
"she had marked him down as a dangerous liberal"
   - acknowledge or celebrate (an important event) with a particular action.
"to mark its fiftieth birthday the charity held a fashion show"
Synonyms: celebrate · observe · recognize · acknowledge · keep · honour · solemnize · pay tribute to · salute · commemorate · remember · memorialize
   - be an indication of (a significant event or stage).
"a series of incidents which marked a new phase in the terrorist campaign"
Synonyms: represent · signify · be an indication of · be a sign of · indicate · herald
   - characterize as having a particular quality or feature.
"the reaction to these developments has been marked by a note of hysteria"
Synonyms: characterize · distinguish · identify · typify · brand · signalize · stamp
   - (of a clock or watch) show (a certain time).
"his watch marked five past eight"

4. (of a teacher or examiner) assess the standard of (written work) by assigning points for proficiency or correct answers.
   - reduce the number of marks awarded to a person or their work.
"teachers must mark down GCSE work containing poor grammar"

5. notice or pay careful attention to.
"he'll leave you, you mark my words!"
Synonyms: take heed of · pay heed to · heed · listen to · take note/notice of · pay attention to · attend to · note · mind · bear in mind · give (a) thought to · take into consideration · take to heart · regard

6. (of a player in a team game) stay close to (an opponent) in order to prevent them getting or passing the ball.
"each central defender marks one attacker"
   - catch (the ball) from a kick of at least ten metres.
"I did well at marking the ball"

Old English mearc, gemerce (noun), mearcian (verb), of Germanic origin; from an Indo-European root shared by Latin margo 'margin'.

mark

[mɑːk]

NOUN
marks (PluralNoun)

1. (until the introduction of the euro in 2002) the basic monetary unit of Germany, equal to 100 pfennig; a Deutschmark.
"Germany spent billions of marks to save the French franc from speculators"

2. a former English and Scottish money of account, equal to thirteen shillings and four pence in the currency of the day.
"Sir William left 500 marks for repairing the road to Cambridge"
   - a denomination of weight for gold and silver, formerly used throughout western Europe and typically equal to 8 ounces (226.8 grams).

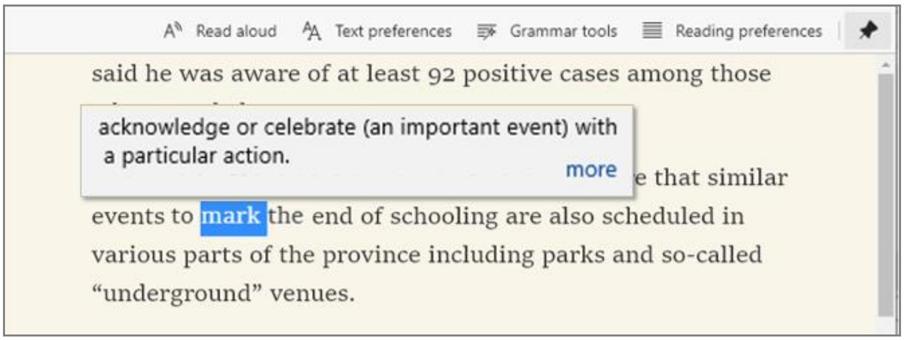
3. the basic monetary unit of Bosnia and Herzegovina, equal to 100 fening.

Old English marc, from Old Norse mork ; probably related to mark1.

Powered by OxfordDictionaries © Oxford University Press

Figure 6: Overview of the complete article after clicking on "more" in Figure 5

A possible solution could be the tagging of word classes, whereby more than half of the lexicographical data shown in Figure 6 could be excluded. The corresponding technology has improved considerably over the past few years, although it is still not completely reliable. As such, tagging could be part of the solution. But even if it were possible to detect the correct word class in the concrete context, the challenge would still be to determine the sense that is relevant in this context (the lexicographical data assigned to the six senses of the verb *mark* represent about 45 percent of the total amount of data in Figure 6). Some of these data may also be superfluous as assistance to reading. But even if they were discarded, there would still be too much data to elegantly fill the default pop-up window activated by touching or clicking on a word in the text.



said he was aware of at least 92 positive cases among those

acknowledge or celebrate (an important event) with a particular action.

more

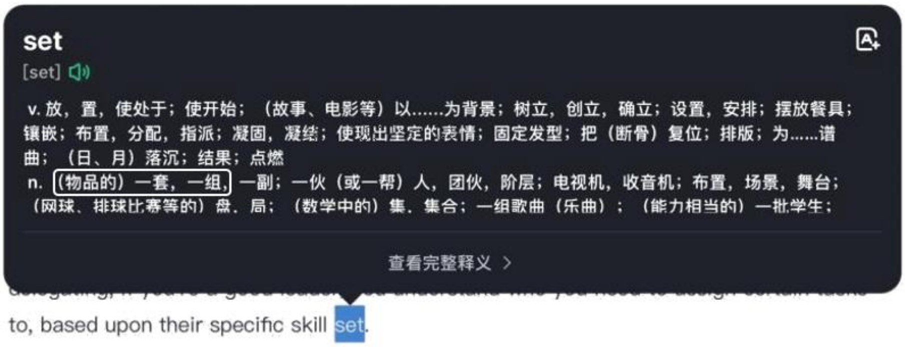
e that similar events to mark the end of schooling are also scheduled in various parts of the province including parks and so-called "underground" venues.

Figure 7: Ideal content of pop-up window to replace the one shown in Figure 5

The ideal solution would be a program that could detect the concrete meaning of any word occurring in a text and, at the same time, only upload the minimum of data required to assist the reader. Figure 7 shows how such a solution could look like, based on the definition marked by the dotted circle in Figure 6. This type of solution is not viable yet, but technology is taken us closer and closer to it. It is now possible to assign a word occurring in a text to a specific sense with a probability of up to 90 percent. The technique makes use of machine learning, big corpora, semantic annotation, and lexicographical databases with a large number of words, senses, and example sentences. Although available, it requires extremely big processing power, and for the time being, only companies like Google can afford to acquire the necessary hardware. To our knowledge, it has not yet been applied to dictionaries integrated into e-readers. But it is under continuous development, and as has happened with other technologies, we can expect it to become cheaper and more efficient within a short span of years. It would probably not be a bad idea that all relevant stakeholders, among them lexicographers and L2 experts, start reflecting and preparing themselves for this Brave New World.

4. Dictionaries integrated into learning apps

The proposed content of the pop-up window shown in Figure 7 is inspired by Huang and Tarp (2021), who suggested a similar solution for an L2-learning app for Chinese learners of English. The two authors based their proposal on a significant difference between learning apps and other tools used to assist the reading of books and Internet texts. The number of texts and words appearing in the latter is literally speaking *infinite*, whereas learning apps include a relatively *limited amount* of texts and words. This difference allows other methods to be applied. The required technology is already available and relatively simple. Instead of artificial intelligence which was a precondition for the proposal in the previous paragraph, the two authors recommend *human-assisted intelligence* based on an interdisciplinary collaboration between language experts, lexicographers, and information engineers.



set
[set]

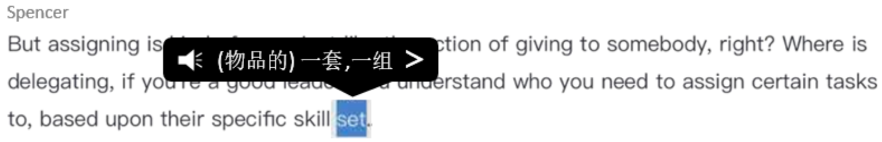
v. 放，置，使处于；使开始；（故事、电影等）以……为背景；树立，创立，确立；设置，安排；摆放餐具；镶嵌；布置，分配，指派；凝固，凝结；使现出坚定的表情；固定发型；把（断骨）复位；排版；为……谱曲；（日、月）落沉；结果；点燃

n. （物品的）一套，一组，一副；一伙（或一帮）人，团伙，阶层；电视机，收音机；布置，场景，舞台；（网球、排球比赛等的）盘，局；（数学中的）集，集合；一组歌曲（乐曲）；（能力相当的）一批学生；

to, based upon their specific skill set.

Figure 8: Pop-up window with *set* in Kaiyan app

To illustrate their proposal, Huang and Tarp (2021) discuss an example from the *Kaiyan OpenLanguage* learning app where a user clicks on the word *set* in the text (see Figure 8). The displayed pop-up window contains a big majority of lexicographical data that are completely irrelevant in the concrete context. More than half of the data belong to the verb *set*, although the word clicked on was a noun. In fact, only the characters inside the white frame relate to the concrete meaning of *set*, i.e. less than 10 percent of the total. This meaning item is not easy to find at first glance. It will probably take several seconds to detect, evaluate, and choose the right meaning of *set*. The large amount of irrelevant data obstructs the information search process. Huang and Tarp (2021) therefore suggest an alternative solution (see Figure 9) and explain how it can be obtained with current technology.



Spencer

But assigning is ction of giving to somebody, right? Where is delegating, if yo erstand who you need to assign certain tasks to, based upon their specific skill set.

(物品的) 一套,一组 >

Figure 9: Ideal content of pop-up window to replace the one shown in Figure 8

The difference between the proposed pop-up window and the one currently used in the *Kaiyan OpenLanguage* app becomes crystal-clear if we compare Figures 8 and 9. The proposal contains an absolute minimum of items to meet the users' needs in the concrete context. Huang and Tarp (2021: 87) explain the underpinning philosophy:

> The main idea is that the pop-up window should only include items that can be justified by the immediate user needs. Thus, it merely consists of a speaker icon, a meaning discriminator followed by two equivalents, and a signifier (>). The central item is the definition (or equivalents) that directly assists understanding of the course text. […] The window also includes a speaker icon to service learners who, as recommended by language didactics, read aloud and may need to listen to some of the words to pronounce them right. Finally, it provides a widely used signifier that affords access to the whole article and is well-known to most netizens.

As can be seen, the proposed default pop-up window breaks with well-established features of the traditional dictionary article, as it does not contain lemma, part of speech, inflectional morphology, other senses, or any other lexicographical data. All these items are considered irrelevant as an immediate response to the specific information need, that is, to understand the text. The proposal is rooted in a millenarian cultural practice.

The traditional dictionary is the result of a long historical development. It started, both in China and Europe, in ancient times when the old scribes copied manuscript works from earlier periods and inserted glosses to explain obsolete and difficult words; see McArthur (1986), Stathi (2006), and Yong and Peng (2008). These glosses were later compiled into glossaries which, over the centuries, developed into the modern dictionary. The process had both advantages and disadvantages and saw at least three important innovations. The first was the invention of the lemma that assigned all grammatical forms of nouns, verbs, and adjectives to a single and generally accepted canonical form. The second was the macrostructure that organises all words treated in the dictionary. In the beginning, it was systematic with the words arranged in the same order as they appeared in a specific book or text. Later, other structuring criteria became dominant. In Europe, the preferred criterion was the alphabet. It started with the first letter of each word, then came the second and the third, and so on, until the macrostructure finally appeared strictly alphabetic. The third invention was the microstructure that became increasingly complex when social and economic development required more and more lexicographical data to be included in the dictionary articles, which also led to a condensed, unnatural description language and the use of codes and abbreviations.

The positive outcome of this long historical process was a lexicographical product (the dictionary) that could be widely consulted in different contexts and not only in connection with a specific text or book. The negative outcome was that the introduction of the lemma and the alphabetic macrostructure required a complex mental process from the users, whereas the microstructure presupposed that the latter developed still better "reference skills".

The solutions proposed in Figures 7 and 9 represent an attempt to build upon the positive aspects and avoid the negative ones detected in the history of lexicography. The default pop-up windows have several advantages:

— The overall design follows the principles of human-centered, or user-centered, design, as recommended by Tarp and Gouws (2020).
— The windows can be activated and used intuitively and do not require special instructions or skills, as recommended by Rundell (2015).
— The response to learners' needs is immediate and represents an example of good communication, as recommended by Norman (2013).
— The lexicographical data are contextualized and provided directly in the context where an information need occurs, as recommended by Tarp and Gouws (2019).
— The windows contain only the required minimum of data and, in this way, avoid negative phenomena such as data and information overload, as recommended by Gouws and Tarp (2017).

All these design features guarantee that the consultation process does not disturb the learners' reading flow and focus on the text, thus creating the optimal conditions for incidental learning as defined above. What is more, the design also provides easy and intuitive access to additional data by clicking on the respective signifiers ("more" and ">"). In this way, the learners can switch to intentional learning whenever they need it for one reason or another. This step allows them to fully benefit from all the positive aspects of the modern dictionary, provided that the presentation of the additional lexicographical data also follows the mentioned principles of user-centered design.

5. Multi-word units of meaning

The translation of the Chinese characters used to explain the meaning of *set* in Figure 9 is a "suite or series, group (of things)". This short definition of *set*, as it appears in the text, is sufficient to make it understandable to the learner. But it is not completely satisfactory. In the specific context, *set* is part of the frequent word combination *skill set*, which represents a so-called extended unit of meaning. In a posthumous article, Sinclair (2010: 37) discusses the lexicographical treatment of this type of multi-word combinations and recommends their lemmatization:

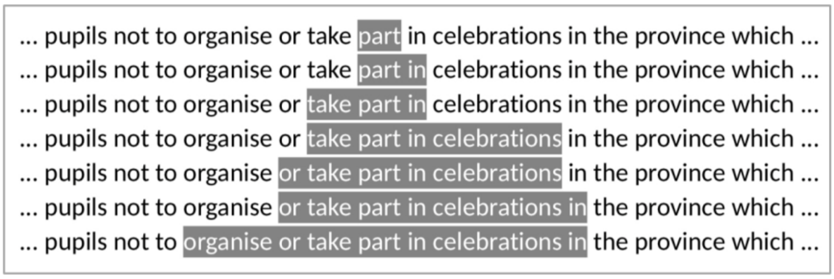
> The evidence from corpora adds up to a strong case for extending the treatment of multi-word units of meaning — a much wider concept than idiom — and giving them the same status as the usual headword.

Sinclair's reflections are also relevant to e-reading tools. Learners cannot be expected to recognise multi-word units of meaning when they meet them in a text. If they do not understand them, they will tend to click on the individual words — in the above case, either *skill* or *set*, or both of them separately. The misinterpretation may derail the consultation process with negative consequences for the reading flow. Designers of e-reading tools should therefore make provisions for this challenge. The tools should be designed to give a lexicographical response that covers the multi-word units of meaning as a whole when users click on one of their component parts. Huang and Tarp (2021) have shown how it could be done in learning apps combining good handicraft and relatively simple programming. But the suggested method presupposes a limited amount of texts and is not an option in e-readers and similar tools where the texts, in theory, are unlimited (see above). Fortunately, current technology makes allowance for another method that is already used to analyse texts. To illustrate how it works, we have taken the following sentence from the article used in Paragraph 3 (Singh 2020):

> Zikalala appealed to parents and pupils not to organise or take part in celebrations in the province which flout current Covid-19 safety protocols and endanger lives.

If readers, especially non-native speakers of English, have problems understanding this sentence, they may click on one or more words, for instance *part*, to get lexicographical assistance. The underlying program then automatically starts exploring the surrounding words to detect extended units of meaning.

Figure 10 is a schematic representation of the process that takes place and only lasts a few nanoseconds. It goes more or less like this: The program starts looking at the first word after *part* to see if there is a recognisable multi-word unit; it then continues with the first word before *part*, the second word after *part*, the second word before *part*, and so on. In the example shown in Figure 10, it examines the six words closest to *part*, but it could be programmed to do more. In this way, it can detect extended units of meaning consisting of two or more words, even if they are separated from each other by a few other words.



... pupils not to organise or take part in celebrations in the province which ...
... pupils not to organise or take part in celebrations in the province which ...
... pupils not to organise or take part in celebrations in the province which ...
... pupils not to organise or take part in celebrations in the province which ...
... pupils not to organise or take part in celebrations in the province which ...
... pupils not to organise or take part in celebrations in the province which ...
... pupils not to organise or take part in celebrations in the province which ...

Figure 10: Schematic representation of technique to detect multi-word units

In the sentence from Times Live, *take part* is a multi-word unit with its own specific meaning (*participate*). Hence, when readers are unaware of this and click on *part*, the described technique makes it possible to give them a lexicographical response to *take part*. The instantaneous and context-adapted response is likely to boost their learning of this and similar multi-word units of meaning. The precondition, however, is that the underlying program already "knows" these units and can react upon them. This underscores the relevance of their lexicographical treatment and lemmatisation as recommended by Sinclair (2010). As mentioned above, the described technique is already available but, as far as we are informed, it has not yet been applied to dictionaries integrated into e-reading tools. Its application does not only depend on information engineers. It also requires that lexicographers compile the high-quality lexicographical databases that allow cutting-edge programming to prosper for the benefit of L2 learners, whether engaging in incidental or intentional learning.

6. Digital writing assistants

In a recent article, Graham (2020: 535) contends that:

> the sciences of reading and writing are too narrowly focused on how to teach either reading or writing and not focused enough on how these two skills can be used to support each other.

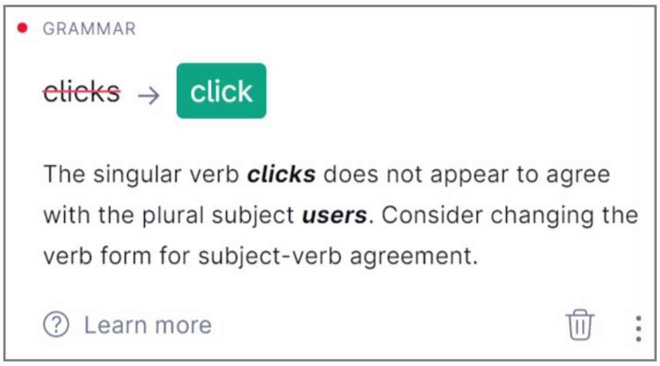
Graham, therefore, recommends that the two sciences become "more fully integrated". From the perspective of lexicography, these statements seem both logical and relevant for the topic under discussion. Just like reading, writing is increasingly performed on digital devices. Together with oral text production, it is through writing that learners activate their L2 vocabulary and train spelling, morphology, and syntax. Similar to what happened in Figures 2 and 3, writing can also be a source of incidental learning. Where the former was human-assisted, incidental learning in connection with writing must be machine-assisted.

Digital writing assistants have some interesting possibilities in this respect, especially those integrated into the text-processing programs learners typically use when they write. When these tools, for instance, are installed on smartphones and tablets and indicate the most likely word terminations, they may stimulate spelling; see, e.g., the one presented by Tarp et al. (2017). When they suggest the most likely words to follow in the sentence, they guide their users into the exciting world of word combinations, among them the ones discussed in the previous paragraph. In this respect, Hanks (2013: 399) distinguishes between *possible* and *probable* combinations and observes that "the number of probable combinations … is rather limited", although "the number of possible combinations may in principle be limitless". Rundell (2018: 6) adds:

> Although corpus analysis enables us to observe the inbuilt predictability of most language output, much of this is far from predictable to a learner …

The desired predictability can be boosted by some of the techniques applied in writing assistants. However, it is not sufficient to look forward based upon the words already typed. It is also necessary to look back on these words to check whether one or more of them has to be changed. This is the advantage of the technology applied in writing assistants like Grammarly, LanguageTool, and ProWritingAid. These tools do no predict the next word in the sentence, but they come up with alerts and suggestions only a few seconds after typing the words. So far, the technology only seems to handle word combinations to a certain extent and, thus, still needs to be improved in this regard. It is, however, increasingly efficient and convincing in other aspects like word choice, spelling, morphosyntax, and punctuation. The suggestions refer to parameters like correctness, clarity, conciseness, and conventions, and include both error correction and text improvement marked with different colors. When users click on the marked words, a small window pops up with a brief explanation and an alternative solution, which they can insert into the text with another click. Figure 11 shows the content of the pop-up window activated after replacing *click* with *clicks* in the previous sentence.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1703 (Article)



● GRAMMAR

~~clicks~~ → click

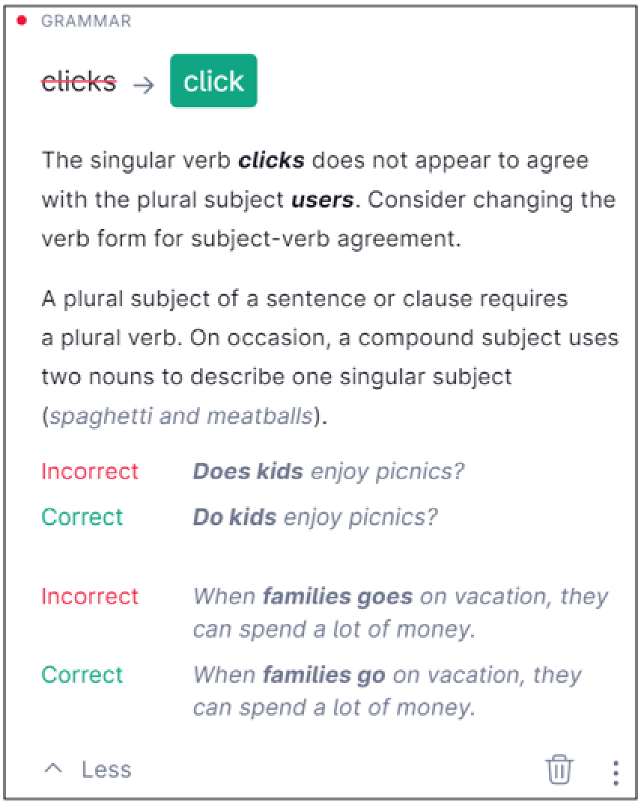
The singular verb ***clicks*** does not appear to agree with the plural subject ***users***. Consider changing the verb form for subject-verb agreement.

Figure 11: Pop-up window activated by clicking on an alert in Grammarly

In his visionary reflections on dictionaries, Sweet (1899: 139) also had some important recommendations concerning the treatment of grammar:

> A thoroughly useful dictionary ought, besides, to give information on various grammatical details, which, though they fall under general rules of grammar, are too numerous or too arbitrary and complicated to be treated of in detail in any but a full reference-grammar: such a dictionary ought to give full information about those grammatical constructions which characterize individual words, and cannot be deduced with certainty and ease from a simple grammatical rule. (Sweet 1899: 139)

The discussion raised by Sweet (1899) has been going on since then. Lexicographers have defended different positions on the relationship between grammar books and dictionaries and have proposed various principles for the inclusion of grammatical data and how to treat them; see Jackson (1985), Mugdan (1984), Cowie (1987, 1989), Herbst (1989), Rundell (1998), Bogaards and Kloot (2001), among many others. Agreement has not been reached, although the general tendency is to introduce still more grammatical data explained in plain language without unnecessary grammatical codes and abbreviations. It is, therefore, interesting to observe how Grammarly seems to share, at least partially, the above vision expressed by Sweet (1899). If the writers click on the signifier "Learn more", they will get immediate access to an extended usage note that explains the relevant grammatical problem in greater detail (see Figure 12). In this way, the general grammatical rule materialises in a mini-rule assigned to an individual word. It can be argued that the solution is not completely individualised and, thus, successful, but this is a case for further analysis and improvement.



~~clicks~~ → click

The singular verb ***clicks*** does not appear to agree with the plural subject ***users***. Consider changing the verb form for subject-verb agreement.

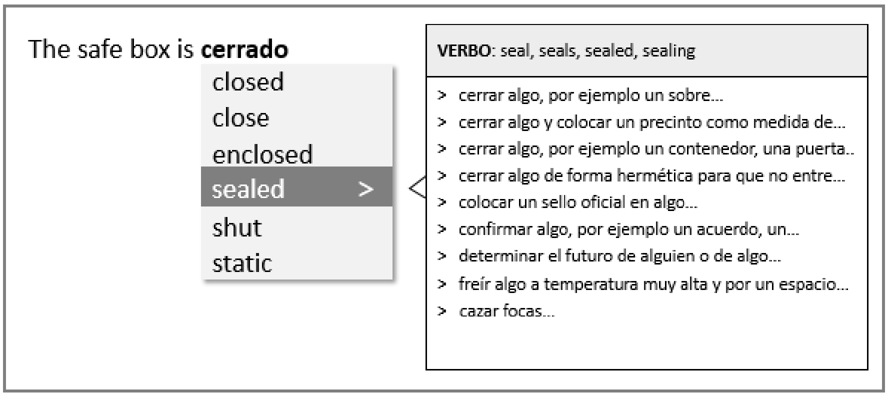
A plural subject of a sentence or clause requires a plural verb. On occasion, a compound subject uses two nouns to describe one singular subject (*spaghetti and meatballs*).

Incorrect	***Does kids*** *enjoy picnics?*
Correct	***Do kids*** *enjoy picnics?*
Incorrect	*When **families goes** on vacation, they can spend a lot of money.*
Correct	*When **families go** on vacation, they can spend a lot of money.*

Figure 12: Grammar rule activated by clicking on "learn more" in Figure 10

The almost instantaneous suggestions and the option to insert the suggested corrections directly in the text with a simple click may pave the way for incidental learning, in so far as the production of a correct text, and not the intent to learn, is the writer's primary goal. Even so, there is undoubtedly a delicate balance between writing flow and provision of information, between focus on the text and focus on the consultation. But when the user consciously decides to click on "Learn more", it is definitely a case of intentional learning.

In Figure 3, we saw a bilingual approach to incidental learning in connection with oral communication when an L2 learner switched to his native language to ask for an L2 word he did not know or remember. Now we will demonstrate something similar in connection with written communication. We will use an example from Write Assistant (see Figure 13) discussed by Fuertes-Olivera and Tarp (2020). A Spanish learner is writing a text in English, and when he lacks an English word to express a specific idea, he types the Spanish word *cerrado* instead. The software is designed to be context-aware and automatically displays several equivalents with the ones that are most likely in the concrete context listed first. This prioritised list has two functions. It can function as a reminder if the learner already knows the word but has just forgotten it. In this case, he can simply click on the appropriate equivalent to introduce it directly into the text. If he does not know one of the equivalents (in this case *sealed*), he can mark it and click on the arrow to activate a pop-up window with short L1 definitions of its various senses written in his mother tongue. If one of these meets his expectations, he can click on *sealed* to insert it into the text. The problem is solved, and he can continue writing. But if he, for one reason or another, wants to know more, he can click on one of the arrows in the pop-up window to access more lexicographical data related to the specific sense, as explained in detail by Fuertes-Olivera and Tarp (2020). If he chooses to do this, it is once more a case of intentional learning. By contrast, if the short definitions in the pop-up window are sufficient to meet his concrete needs, it may allow the learning of one of the senses of *seal*. This learning is, by definition, incidental, as long as the writer's direct and immediate intention is not to learn but to express something through a written text.



The safe box is **cerrado**

VERBO: seal, seals, sealed, sealing

- > cerrar algo, por ejemplo un sobre...
- > cerrar algo y colocar un precinto como medida de...
- > cerrar algo, por ejemplo un contenedor, una puerta..
- > cerrar algo de forma hermética para que no entre...
- > colocar un sello oficial en algo...
- > confirmar algo, por ejemplo un acuerdo, un...
- > determinar el futuro de alguien o de algo...
- > freír algo a temperatura muy alta y por un espacio...
- > cazar focas...

Figure 13: Prioritised equivalents and pop-up window in Write Assistant

7. Conclusions

From a lexicographical perspective, incidental learning presupposes instantaneous, contextualized, and unobtrusive assistance with an absolute minimum of lexicographical data, whereas intentional learning requires easy access to relevant additional data. The discussion above shows that it is far more complicated to create the conditions for the former than the latter. In both cases, the lexicographical data must be high quality and presented in a user-centered design that allows intuitive use. But incidental learning, as defined in Paragraph 2, also implies that the software has been trained to exclusively present the data needed in each concrete case and discard all other data. As we have seen, the techniques and technologies required are already available to a certain extent, albeit insufficiently applied to lexicography. They include artificial intelligence, human-assisted intelligence, and cutting-edge programming in general. If further developed and fully integrated into lexicography, they herald a major technological breakthrough with the advent of context-aware lexicographical products, that is, a prototype of intelligent dictionaries.

The creation and quality of future context-aware lexicographical products do not only depend on the successful application of cutting-edge technologies. It also requires that lexicographers reconsider part of their discipline. It implies, among other things, refinement of the lexicographical databases that store the pertinent data and improved design of the user interfaces that present the relevant data to the target users. As Zhang (2019) rightly asserts, the current media convergence age invites lexicographers to take an innovative approach to the compilation and publication of dictionaries and, could it be added, lexicographical products in general. The challenge is *both* to integrate various media into digital dictionaries, *and* to integrate lexicography into different types of digital devices.

It may seem that some of the issues discussed in this paper go beyond traditional lexicography or, at least, belong to the borderland between lexicography and other fields of endeavour. That may be so. In any case, a better interpretation would be that they represent virgin land which the millennial discipline needs to cultivate if it wants to meet contemporary challenges. It requires a successful symbiosis of tradition and disruption; see Tarp (2019). The suggestions to improve current e-reading tools, writing assistants, and learning apps intend to remove some of the obstacles to incidental L2 learning in connection with the reception and production of written texts. It is up to future research to determine whether or not, and to what degree, these suggestions work in practice. Until then, they remain a digital information-assisted possibility of incidental learning.

References

A. Writing assistants

Grammarly. https://www.grammarly.com/.

LanguageTool. https://languagetool.org**.**

ProWritingAid. https://prowritingaid.com/.

Write Assistant. https://writeassistant.com/.

B. Other literature

Bereiter, C. and M. Scardamalia. 1989. Intentional Learning as a Goal of Instruction. Resnick, L.B. (Ed.). 1989. *Knowing, Learning, and Instruction: Essays in Honor of Robert Glaser*: 361-392. Hillsdale: Lawrence Erlbaum Associates.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1703 (Article)

Bogaards, P. and W. van der Kloot. 2001. The Use of Grammatical Information in Learners' Dictionaries. *International Journal of Lexicography* 14(2): 97-121.

Bothma, T.J.D. and R.H. Gouws. 2020. e-Dictionaries in a Network of Information Tools in the e-Environment. *Lexikos* 30: 29-56.

Bothma, T.J.D. and D.J. Prinsloo. 2013. Automated Dictionary Consultation for Text Reception: A Critical Evaluation of Lexicographic Guidance in Linked Kindle e-Dictionaries**.** *Lexikographica* 29(1): 165-198.

Brown, R., R. Waring and S. Donkaewbua. 2008. Incidental Vocabulary Acquisition from Reading, Reading-while-listening, and Listening to Stories. *Reading in a Foreign Language* 20(2): 136-163.

Cowie, A.P. 1987. Syntax, the Dictionary and the Learner's Communicative Needs. 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*: 183-92. Tübingen: Niemeyer.

Cowie, A.P. 1989. Information on Syntactic Construction in the General Monolingual Dictionary. Hausmann, F.J., O. Reichmann, H.E. Wiegand and L. Zgusta (Eds.). 1989. *Wörterbücher, Dictionaries, Dictionnaires. An International Encyclopedia of Lexicography. Volume 1*: 588-592. Berlin/ New York: Walter de Gruyter.

Fuertes-Olivera, P.A. and S. Tarp. 2014. *Theory and Practice of Specialised Online Dictionaries: Lexicography versus Terminography.* Berlin/Boston: De Gruyter**.**

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

Gouws, R.H. and S. Tarp. 2017. Information Overload and Data Overload in Lexicography. *International Journal of Lexicography* 30(4): 389-415.

Graham, S. 2020. The Sciences of Reading and Writing Must Become More Fully Integrated. *Reading Research Quarterly* 55(S1): 535-544.

Hanks, P. 2013. *Lexical Analysis: Norms and Exploitations*. Cambridge, Mass./London: MIT Press.

Herbst, T. 1989. Grammar in Dictionaries. Tickoo, M.L. (Ed.). 1989. *Learner's Dictionaries: The State of the Art*: 94-111. Singapore: SEAMO Regional Language Centre.

Huang, F. and S. Tarp. 2021. Dictionaries Integrated into English Learning Apps: Critical Comments and Suggestions for Improvements. *Lexikos* 31: 68-92.

Hulstijn, J.H. 1989. Implicit and Incidental Second Language Learning: Experiments in the Processing of Natural and Partly Artificial Input. Dechert, H.W. and M. Raupach (Eds.). 1989. *Interlingual Processes*: 49-73. Tübingen: Gunter Narr.

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.

Jackson, H. 1985. Grammar in the Dictionary. Ilson, R. (Ed.). 1985. *Dictionaries, Lexicography and Language Learning*: 53-59. Oxford: Pergamon Press.

Krashen, S. 1989. We Acquire Vocabulary and Spelling by Reading: Additional Evidence for the Input Hypothesis. *Modern Language Journal* 73*:* 440-464.

Laufer, B. and J. Hulstijn. 2001. Incidental Vocabulary Acquisition in a Second Language: The Construct of Task-induced Involvement. *Applied Linguistics* 22: 1-26.

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.

Lessing, G.E. 1747. *Der Junge Gelehrte*. Stuttgart: Reclam 1965.

McArthur, T. 1986. *Worlds of Reference. Lexicography, Learning and Language from the Clay Tablet to the Computer.* Cambridge: Cambridge University Press.

Mugdan, J. 1984. Grammatik im Wörterbuch: Wortbildung. Wiegand, H.E. (Ed.). 1984. *Studien zur neuhochdeutschen Lexikographie IV*: 237-308. Hildesheim/Zürich/New York: Olms.

Nagy, W.E., P.A. Herman and R.C. Anderson. 1985. Learning Words from Context. *Reading Research Quarterly* 20(2): 233-253.

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

Rundell, M. 1998. Recent Trends in English Pedagogical Lexicography. *International Journal of Lexicography* 11(4): 315-342.

Rundell, M. 2015. Review Article: Shigeru Yamada. Oxford Guide to the Practical Usage of English Monolingual Learners' Dictionaries: Effective Ways of Teaching Dictionary Use in the English Class. *Kernerman Dictionary News* 23: 26-27.

Rundell, M. 2018. Searching for Extended Units of Meaning — and What To Do When You Find Them. *Lexicography — Journal of Asialex* 5(1): 5-21.

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.

Sinclair, J.M. 2010. Defining the Definiendum. De Schryver, G.-M. (Ed.). 2010. *A Way with Words: Recent Advances in Lexical Theory and Analysis*. *A Festschrift for Patrick Hanks*: 37-47. Kampala: Menha Publishers.

Singh, O. 2020. KZN Premier Lays Down the Law Against Those Attending 'Superspreader' Events. *Times Live*, 8 December 2020.

Stathi, E. 2006. Greek Lexicography, Classical. Brown, K. (Ed.). 2006. *Encyclopedia of Language and Linguistics. Vol. 5*: 145-146. Second Edition. Oxford: Elsevier.

Sweet, H. 1899. *The Practical Study of Languages: A Guide for Teachers and Learners*. London: J.M. Dent. Reprinted in 1964 by Oxford University Press.

Tarp, S. 2008*. Lexicography in the Borderland between Knowledge and Non-knowledge. General Lexicographical Theory with Particular Focus on Learner's Lexicography.* Tübingen: Max Niemeyer.

Tarp, S. 2019. Connecting the Dots: Tradition and Disruption in Lexicography. *Lexikos* 29: 224-249.

Tarp, S., K. Fisker and P. Sepstrup. 2017. L2 Writing Assistants and Context-Aware Dictionaries: New Challenges to Lexicography. *Lexikos* 27: 494-521.

Tarp, S. and R.H. Gouws. 2019. Lexicographical Contextualization and Personalization: A New Perspective. *Lexikos* 29: 250-268.

Tarp, S. and R.H. Gouws. 2020: Reference Skills or Human-Centered Design: Towards a New Lexicographical Culture. *Lexikos* 30: 470-498.

Wiegand, H.E. 2013. Lexikographie und Angewandte Linguistik. *Zeitschrift für angewandte Linguistik* 58(1): 13-39.

Yong, H. and J. Peng. 2008. *Chinese Lexicography. A History from 1046 BC to AD 1911.* Oxford: Oxford University Press.

Zhang, Y. 2019. On the Innovation of Dictionary Compilation and Publication in the Context of Media Convergence. *Research on Language Strategies* 4(6): 79-89.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1704 (Article)

Towards an Evolutional Chain of English Dictionary Paradigms from the Linguistic Perspective

Heming Yong, *Guangdong University of Finance, Guangzhou, People's Republic of China (hmyong818@sina.com)*

Abstract: This paper aims to unfold, by tracing the evolutional thread of English dictionaries from their earliest roots to present state from the linguistic perspective, a coherent and complete picture of how English dictionary making develops from its archetype to the prescriptive, the historical, the descriptive and finally to the cognitive form. It builds up an integrated chain of English dictionary paradigms and demonstrates how English lexicography develops into its modern form through inheritance, innovation and self-perfection.

Keywords: ENGLISH LEXICOGRAPHY, DICTIONARY PARADIGMS, ARCHETYPE, LATIN TRADITION, PRESCRIPTIVISM, DESCRIPTIVISM, DIACHRONISM, COGNITIVISM

Opsomming: Op weg na 'n evolusionêre reeks Engelse woordeboekparadigmas vanuit 'n linguistiese perspektief. In hierdie artikel word gepoog om 'n koherente en volledige prentjie te skets van hoe Engelse woordeboekmaak ontwikkel het vanaf argetipe tot preskriptiewe, historiese, deskriptiewe en uiteindelik kognitiewe vorm deur die evolusionêre "draad" van Engelse woordeboeke vanaf hul oorsprong tot die huidige stand vanuit 'n linguistiese perspektief na te spoor. 'n Geïntegreerde reeks Engelse woordeboekparadigmas neem vorm aan en daar word gedemonstreer hoe die Engelse leksikografie deur nalatenskap, vernuwing en selfvervolmaking tot die moderne vorm ontwikkel het.

Sleutelwoorde: ENGELSE LEKSIKOGRAFIE, WOORDEBOEKPARADIGMAS, ARGETIPE, LATYNSE TRADISIE, PRESKRIPTIWITEIT, DESKRIPTIWITEIT, DIACHRONISME, KOGNITIVISME

0. Introduction

English dictionaries can be traced back to the glossaries in the 7th and 8th centuries, and the theoretical roots of English lexicography grew out of Latin dictionary traditions and prescriptivism. Signs of prescriptivism were already discernible in early English dictionary compilation. Latin lexicographical traditions exerted gradual and yet profound influence upon prescriptivism, which became firmly established with the publication of Samuel Johnson's (1709–1784) *A Dictionary of the English Language* (1755).

Towards the late part of the 18th century, historical comparative linguis-

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1704 (Article)

tics came into vogue in the linguistic circles of Europe. Through its evolution in the 19th and the early 20th century, a set of historical linguistic principles, along with comparative methods and internal reconstruction and explorations of word origins from phonological, morphological and semantic aspects, evolved into the historical dictionary paradigm, which was amply taken advantage of in *The Oxford English Dictionary* (1884–1933).

Language description was widely recognized as the mainstream approach of the 20th-century linguistic research, and descriptivism triggered off revolutionary changes in notions, principles, methodological and theoretical formulation directly related to dictionary making. Compilers started to adapt themselves to changes in the trends of linguistic study and turn their dictionaries into language recorders and describers rather than authorities and arbitrators. Descriptivism became an established practice in Philip Babcock Gove's (1902–1972) *Webster's Third New International Dictionary of the English Language,* Unabridged (1961).

Dictionary compilation used to be separated from dictionary use and language cognition, and dictionary compilation and research are bound to be seriously defective without taking the user perspective into consideration. *Longman Dictionary of Contemporary English* (1978) ushered in a new era of cognitivism characterized by unique focuses on users and seamless integration of dictionary design and dictionary use, dictionary function and language cognition, and dictionary making and electronic technology, highlighted by *The WordNet* online.

The concept of "paradigm" was introduced into lexicographical studies only decades ago, referring to a model, pattern or a set of principles for dictionary design, compilation and research. This paper attempts to explore the historical trajectory of English dictionary paradigms from the linguistic perspective with a view to revealing the interactive mechanisms and the historical inheritance between the evolution of English dictionary paradigms and the progress of linguistic theories, particularly modern linguistics.

1. The archetype of English dictionary paradigms

A general survey of the origins of world lexicographical culture manifests two discernible sources of development. One is the collection and accumulation of annotations and notes left on the margins and between lines of ancient classic works by the so-called authorities or social elites, such as monks, missionaries, priests and schoolmasters, and the other is the glossaries compiled collectively by people with expertise to meet special needs of religious preaching, literacy education, national assimilation, and military occupation.

These glossaries and vocabularies are found in ancient Chinese, Latin, Greek and Sanskrit. They were compiled, revised, enlarged or augmented over time into larger and more comprehensive volumes. Early works were made either monolingually or bilingually from annotations and explanatory notes collected from various classic works. On rare occasions they might be collec-

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1704 (Article)

tions of such annotations and notes from only one work, and their entries are arranged in the sequence of their appearance in the text, rather than on the alphabetical or thematic basis. The historical literature demonstrates that English dictionary paradigms originated from explanatory notes and textual researches in the classic works of the Old English period.

The practice of providing annotations in the history of English lexicographical culture can be traced back to the prehistoric Celtic and Germanic languages. Those pioneers who provided such marginal notes or glosses to words, particularly rarely used hard words, were priests and then schoolmasters (Murray 1900; Krache 1975). "And these beginnings themselves, although the English Dictionary of to-day is lineally developed from them, were neither Dictionaries, nor even English" (Murray 1900: 7). However, they turn out to be extremely valuable to modern philologists, as they are a record of words and expressions that could appear in no other sources than Old English, Old Irish and old Germanic languages.

For the convenience of preaching and teaching scriptures, smart monks and schoolmasters started to collect the explanatory notes from between the lines and margins of the text into "glossariums" or glossaries. This is the first distant source of English lexicography. Another early source is the classified glossaries or vocabulary lists that were made for the purpose of Latin learning and teaching and for the convenience of memorizing Latin words and, in most cases, provide explanations of word meanings in English or dialectal vernaculars. They signify the inception of English dictionary paradigms.

The beginnings of English dictionaries "lie far back in times almost prehistoric" (Murray 1900: 7), and no textual research can justify the exact dates of the appearance of the earliest glossaries. However, the fountain-heads of English lexicography can undoubtedly be traced back to the early Anglo-Saxon times, "to a time somewhere between 600 and 700 A.D., and probably to an age not long posterior to the introduction of Christianity in the south of England at the end of the sixth century" (Murray 1900: 13). At the turn of the 7th and 8th century, collections of Latin hard words explained in simpler Latin or Old English began to appear, and the earliest extant one, the *Leiden Glossary*, which was made c. 800 in the Abbey of Saint Gall on the basis of earlier Anglo-Saxon exemplars, comes down to us in the form of manuscripts copied in the 9th century (Murray 1900: 12-13; Green 1996: 55).

The *Leiden Glossary* contains 48 *glossae collectae* (or chapters), and each chapter is prefixed with the title of the text from which the lemma are taken, and the lemmata are arranged in the sequence of their appearance in the text. "Most of the glosses are in Latin, though 250 of them are in Old English." They not only "explain terms from texts used in the classroom", thus a "record of their classroom teaching", but give evidence of the impressive holdings of the Canterbury library (none of which remains) and the reading interests of Anglo-Saxon churchmen" as well (Wikipedia, Leiden Glossary entry; Sauer 2009: 34).

It can be assumed that this glossary was of valuable help to those who

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1704 (Article)

learned how to read and spell when used alongside with the texts but would be substantially discounted in value when used separately. It is not hard to imagine the great inconvenience in looking up lemmata. Users will have to go through the whole glossary in order to find one lemma, and sometimes, repeated searches will have to be made. It is perhaps worth mentioning that the last part of the glossary is a short collection of ancient vocabulary, animal names and terms for other things. The significance of this last part lies in that it is itself a miscellany of words and terms, that it is encyclopedic in nature, and more importantly, that it is the archetype of most of the later glossaries and vocabularies in terms of lemmata collection and their thematic classification.

Alphabetization can be traced back to a glossary of difficult words in Homer's works compiled by Zenodotus (c. 325–c. 234 B.C.) (Collison 1982: 26), but it came into use in English glossaries at quite a late time. The use of the *Leiden Glossary* as a reference tool would have been substantially facilitated if its lemmata had been put into alphabetical order. So, when reproductions of this and other glossaries were made later with augmentations from other sources, all the lemmata beginning with the same letter were extracted and listed together so that the first-letter order was implemented. Improvements were found in *The Epinal Glossary*, *The Corpus Glossary*, and *The Erfurt Glossary*. By about 725, when *The Corpus Glossary* was compiled, the alphabetical principle was advanced to second-letter order (Wells 1973: 13), so that the first 95 lemmata began in Ab- and what followed began in Ac-, and so on. The alphabetical principle began to take precedence over the thematic principle. In an anonymous 10th-century glossary in the British Museum, the alphabetization of some lemma was carried as far as the third letter.

Just as almost all lexicographical cultures in the world originate from explanations of hard words and expressions, so English lexicography started from the practice of providing explanations for hard Latin words and expressions, primarily the annotations of hard Latin words by simple or easier Latin words, and occasionally by Old English vocabulary. Consequently, the frequency of Old English words in early glossaries is extremely low, e.g. only 10% of the word count in *The Epinal Glossary*. But that rose to a considerable level when *The Corpus Glossary* came into existence. Subsequently, no matter how lemmata were arranged, Latin gradually gave way to Old English as the defining language. By the 11th century, almost every Latin lemma was provided with one explanatory English equivalent, and even several equivalents in some cases. Those are the earliest beginnings of Latin–English lexicography, marking the emergence of the English bilingual dictionary paradigm and paving way for the flourishing of Latin–English and English–Latin lexicography.

The lemmata in early glossaries are all Latin, with Latin definitions and explanations. Only when there are no proper Latin words are words and expressions of Old English and vernaculars used for defining and explaining Latin lemmata. By nature, they are merely simple monolingual vocabulary lists, with definitions or explanations only occasionally written in languages other than Latin. English bilingual lexicography started to reach its first climax with the rise of the Renaissance in the 12th century, so the development of English monolingual lexicography was hampered to some extent.

However, thanks to the developed paradigm and referential values established by such English bilingual dictionaries as *Thesaurus Linguae Romanae et Britannicae* (1565) by Thomas Cooper (c.1517–1594) and *Dictionarium Linguae Latinae et Anglicanaev* (1587) by Thomas Thomas (1553–1588), the monolingual *A Table Alphabeticall*, which was written by Robert Cawdrey (c.1538–c.1640), manifests well-conceived configurations concerning its macrostructure and microstructure at its appearance in 1604. It is the first collection of its kind and is recognized as the first English monolingual dictionary with its structural organization resembling that of modern dictionaries to a greater extent than ever before. That can be considered the origin of the paradigm for English monolingual dictionaries.

An overview of ancient dictionary development from the global perspective shows that, no matter how long and under what background their compilation takes, their data sources are no other than collections of glosses out of the ancient manuscripts of religious scriptures and classic works and their combination into different word lists. Explanatory notes or annotations are generally found above the words, between the lines or along the margins in ancient classic works and scriptures.

The beginnings of English dictionary making demonstrate similar patterns and paths, which were inherited by Cawdrey in his compilation of *A Table Alphabeticall*. Although he paid considerable attention to new words, inkhorn terms and bigger issues such as the nature of language, Cawdrey's initial interest was still in explaining "hard vsual English wordes" and fossilizing their spellings "for the benefit and helpe of Ladies, Gentlewomen, or other unskillful persons". Functionally, *A Table Alphabeticall* is didactic rather than descriptive.

The first edition of *A Table Alphabeticall* listed 2652 headwords. Each entry is generally no more than one line, with very simple definitions, usually written in single words or synonyms and synonymous expressions. Indications of word origins are given by means of abbreviations, such as [fr] for French (e.g. [fr] cancell, to vndoe, deface, crosse out, or teare) and (g) for Greek (e.g. throne, (g) a kings seate, or chaire of estate). In rare cases, indications of sense relations are even given, such as the use of "k" (i.e. a kind of) to suggest hyponymy (e.g. lethargie, (g) (k) a drowsie and forgetfull disease). As Cawdrey's intention is to provide meanings of hard words and codify their spellings, *A Table Alphabeticall* has a strong flavor of linguistic purism and prescriptivism. However, it signifies an important transition of English dictionary paradigm from glossaries and vocabularies to dictionaries in a somewhat modern sense and triggers off sparks of prescriptivism in English lexicography.

2. The prescriptive paradigm of English lexicography

Linguistics can be divided from a functional perspective into traditional linguistics (typically traditional grammar), descriptive linguistics (typically descriptive grammar), encoding linguistics (typically transformational generative grammar), decoding linguistics (typically systemic functional grammar), etc. This functional approach is of more lexicographical significance in providing rich lexicographical implications. English lexicography has its theoretical beginnings in the prescriptivism of traditional linguistics.

Prescriptivism is established on the assumption that like all other things, language use should be conducted in the "correct" way. Classic linguists claim that rules should be made for the best or the "most correct" use of language. Prescriptive grammar is based on their views of the best language usage rather than on the description of actual language use. It adopts such criteria as purity, logic, historical and literary superiority to pass judgment upon the best language use and make norms for it. Any deviation from or violation of language norms is treated as language decay and corruption, and should be avoided, purified and put right in the light of logic and literary supremacy, just to prevent linguistic pollution and decay.

Research by British scholars in the 1980s show that English prescriptivism goes back to one of the Middle English varieties called Chancery English, the official written English that developed at the Court of Chancery, was used in administrative documents instead of French after about 1430 and eventually became the base for spelling regularization (e.g. gaf/gave rather than yaf, such rather than swich, theyre/their rather than hir). Chancery English marked the beginnings of a national standard of English spelling, vocabulary and grammar. By the 15th century, printing technology came from China to Europe and was introduced to Britain by William Caxton (c.1415/1422–1492) in 1476. The grammar and spellings Caxton adopted are mainly derived from Chancery English and became the foundation for purifying and codifying English spellings, which are the earliest traces of prescriptivism in the evolution of the English language.

Prescriptivism presented itself in the Old English period in the form of linguistic purism (also known as linguistic protectionism). This linguistic ideology assumes that decay or corruption will take place in a language when deviations occur from ideal language norms, or contacts occur between two languages so that linguistic similarities are produced, and that whatever modifications take place will have to be prevented, purified, and remedied. This fad of linguistic purism was widespread during the reign of Louis VIII le Lion (1187–1226) in France and is still observable in the reform of the writing systems characterized by lexical, orthographic, morphological, syntactic, and phonetic purism.

Linguistic purism is often labeled as "conservative", but it is often accepted as part of language policies of those governments that intend to conduct linguistic reforms, for it demonstrates innovativeness in the formulation of linguistic standards. Modern linguists tend to adopt a critical attitude towards the prescriptive approach to language and emphasize the importance of describing

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1704 (Article)

the actual use of language and the necessity of recognizing the social variations of language in explaining language use. Over the past three decades, however, interest has been resumed in objectively reassessing prescriptivism from the socio-cultural dimensions (see Milroy and Milroy 1985; Bartsch 1987). Modern linguists have started to clear up misunderstandings and attempted to identify the positive effects of prescriptivism upon language study.

Modern linguistics shows that the right use of language does not merely mean grammatical correctness or compliance to the norms and standards followed by the majority of well-educated members in a speech community. There is much more to that. "Generally, notions of correctness are not developed for their own sake, but are developed and employed only when they are really necessary" (Bartsch 1987: 10). Bartsch (1987) distinguishes six types of correctness in language: correctness of the basic means of expression, correctness of lexical items, correctness of syntactic form, correctness of texts, semantic correctness and pragmatic correctness. The former three types fall into the formal category, and the latter fall into the functional category. Linguistic correctness has traditionally paid almost exclusive attention to the formal aspects.

The formal-category prescriptivism only flickered in the early English dictionary compilation and did not become a constant principle that ran through the whole dictionary making. No systematic methodology was formed in that process, though compilers were, to more or less extent, working under the influence of the strong prescriptive flavor stemming from Latin grammar. It was not until Samuel Johnson published *The Plan of an English Dictionary* in 1747 that the prescriptive principles for English dictionary making were systematically identified and fully expounded. By 1755, when Johnson's Dictionary met its readership, such principles were firmly established and continued to dominate English dictionary making for nearly 200 years.

Prescriptivism stems from Latin grammar and has been exerting influence upon language education, textbook writing, and dictionary compilation for hundreds of years. Prior to Randolph Quirk et al.'s publication of *A Comprehensive Grammar of the English Language* (1985), almost all books of English grammar were prescriptive in nature. They are still popular to some extent with English learners, particularly with non-English speaking learners. It is unrealistic to get rid of the influence of prescriptivism overnight, and from the perspective of dictionary making, it is inevitable that all dictionaries are to certain extent infused with prescriptive coloring in their making, which is not merely restricted to spelling, because dictionary users tend to regard dictionaries as authorities in their consultation. That explains why prescriptivism was widely accepted and became the dominating principle after its introduction into English dictionary making.

3. The historical paradigm of English lexicography

The historical paradigm of English lexicography is derived from studies of word origins, with its most distant source being traced to the Roman philologist Lucius Aelius Stilo Praeconinus of Lanuvium (152–74 B.C.). "Influenced by the Stoic philosophy, Praeconinus was interested in grammar and etymology, writing numerous articles on these subjects and eventually producing a glossary" (Collison 1982: 27). In 43 B.C., Marcus Terentius Varro (116–27 B.C), "the first important Roman grammarian" and Praeconinus' pupil, published *De lingua Latina*, which "comprised twenty-five books", with separate sections discussing "the origin of words" and "the derivation of words from other words", but "his etymological conclusions were rather more inspired than logically argued" (Collison 1982: 27). In about 430, an Alexandrian teacher called Orion compiled an etymological dictionary, which set an example for several later compilations of a similar kind.

By the 7th century, Isidorus Hispalensis (also known as St. Isidore of Seville, c.560–636), a Spanish scholar, began to compile a noteworthy encyclopedic dictionary — *Originum; seu Etymologiarum libri XX* (twenty books of origins or etymologies), "a book designed as a wide-ranging *vade mecum* by which the newly converted people of Spain might gain access to every aspect of their new, Catholic faith" (Green 1996: 48). The first, among the "twenty books", is in fact an etymological dictionary with alphabetically-arranged headwords. Though it contains many errors and mistakes, many of its elements, particularly its efforts in explorations of word origins, were incorporated into the dictionaries of later years, such as *Catholicon Anglicum*, Hugo's *Derivationes*, Richard *Huloet's Abecedarium Anglico Latinum* (1552), to varying degrees. In 847, Harbanus Maurus (c.776–856) compiled, with numerous adoptions from Isidore's "twenty books", *Opus de universo; sive, De sermonum proprietate*, with one volume devoted to etymologies, a glossary written in much the same style as Isidore's first book.

In the mid-9th century, a glossary of an encyclopedic nature *Etymologicum genuinum* was compiled in Greek, and its author is assumed to be a respectable Greek scholar called Photius (c.825–886). This work itself, again with heavy absorptions and adaptations from his predecessors Herodian, George Choeroboscus, Methodius, Orion, and Theognostus, became a source of borrowings by numerous other works, such as *Etymologicum Magnum*, *Etymologicum Gudianum*, and *Etymologicum Symeonis*. Photius is recognized as the father of Greek etymology. By the 10th century, *Sanas Cormaic* (or *Sanas Chormaic*, also known as Cormac's Glossary) appeared, ascribable to *Cormac úa Cuilennáin* (?–908), an early Irish glossary of over 1400 Irish words with etymologies as well as synonymous explanations or definitions written in simple Irish or Latin. In addition to its observations in old Irish words and expressions, it is a good record of early words of Irish origin and early studies in Irish etymologies.

In the Western world, the practice of providing etymological information in a dictionary started in the middle of the 17th century when Thomas Blount (1618–1679) published *Glossographia* (1656). Blount is one of the earliest lexicographers who attempted to provide etymological information in a systematic fashion.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1704 (Article)

This practice continued in Nathaniel Bailey's (?–1742) *The Universal Etymological English Dictionary* (1721), which "was the first English dictionary to treat etymology with consistent purpose and seriousness" and "is credited with having established etymology as 'one of the requisites of any reputable dictionary'" (Landau 1989: 99). "Bailey listed not only the immediate source of the English word (etymon), but often earlier forms in other languages ... then a novelty". He "was working a century before the great advances in Germanic philology", so it is not surprising that "many of his etymologies appear wildly speculative from our vantage point" (Landau 1989: 46).

Lexicographers' etymological explorations created necessary conditions for the establishment of historical comparative linguistics, which in turn laid a theoretical foundation needed for the making of historical dictionaries. Historical comparative linguistics finds its earliest traces in Rasmus Kristian Rask's (1787–1832) pioneering work of the 19th century, which brought forth two of his major publications — *Introduction to the Grammar of the Icelandic and other Ancient Northern Languages* (1811) and *Anglo-Saxon Grammar* (1817). His works catalyzed the sprouts of comparative Indo-European grammar and clearly delineated relations of origins between Indo-European languages.

Rask was highly cognizant of the primary importance of phonetic laws to the identification of cognate relations and grammatical homogeneity to the persuasiveness of their verification. The core of modern approaches of comparative linguistics stems from Rask's innovative work. The 19th-century accomplishments in philology and in theoretical and practical lexicography helped to achieve complementarity when philologists devoted themselves to dictionary making, which caused fundamental changes in the calibre of dictionary compilers and marked the end of dictionary making by amateurs. Lexicographical professionals came to realize fully what an ideal dictionary should contain and what it should provide for its users. Dictionary users started to look at dictionaries and their making with critical eyes, and their valuable feedback in turn helped to heighten the standards for dictionaries and dictionary compilation. Historical comparative linguistics was by and large accepted as the mainstream of linguistic inquiries in Europe in the middle and late part of the 19th century. That stimulated academic interest in seeking for the origins of words and their languages, and the best way of achieving this end was certainly applying the historical linguistic principles to dictionary making.

In 1857, The Philological Society of London began its discussions about the feasibility of compiling a dictionary on historical principles, but it was not until 1884 that its unbound fascicles began to appear, under the name of *A New English Dictionary on Historical Principles*; Founded Mainly on the Materials Collected by The Philological Society, unofficially renamed OED in 1895. The full dictionary was republished in ten bound volumes in 1928, and five years later, the title OED fully replaced the former name in all occurrences in its reprinting as twelve volumes with a one-volume supplement. This magnificent dictionary received unprecedentedly wide and enthusiastic acclamation and produced a

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1704 (Article)

long range of derivative dictionaries. A discernable thread can be identified of the inception, evolution and final establishment of the historical dictionary paradigm going from Blount to Bailey, and eventually from the *Deutsches Wörterbuch* to OED.

The historical paradigm for English lexicography extracts its theoretical underpinnings from historical comparative linguistics and historical linguistics and adopts historical principles and comparative approaches as its basic methodology. Focus is laid on the evolution and the representation of words of the same language source over different periods of time with a view to reconstructing the pronunciation, spelling, morphology, syntax and sense relations of words from the perspective of language development and exploring the evolutional traces of words over time and diachronic relatedness of linguistic variations in the light of historical literature and linguistic data. In practice, compilers endeavor to find out about the evolutional attributes and laws for word spellings and meanings and seek for the origins and evolutional patterns of word forms, sounds and meanings from phonological, morphological, syntactic, etymological and dialectal dimensions on the basis of diachronic data and grammatical relations.

The aim of the OED, as indicated on its website, is "to present in alphabetical series the words that have formed the English vocabulary from the time of the earliest records [ca. AD740] down to the present day, with all the relevant facts concerning their form, sense-history, pronunciation, and etymology. It embraces not only the standard language of literature and conversation, whether current at the moment, or obsolete, or archaic, but also the main technical vocabulary, and a large measure of dialectal usage and slang" so as to achieve the purpose of overcoming the seven "principal shortcomings" of contemporary dictionaries identified by Richard Chevenix Trench (1807–1886) (1857). Obviously, those "shortcomings of contemporary dictionaries" are all related in some way to word histories, and therefore are also principal shortcomings in the treatment of etymologies in dictionary making. An adequate awareness of the defects of etymological treatment in contemporary English dictionaries ensured the consistent, comprehensive, systematic and scientific implementation of historical principles in the making of OED, which signifies the firm establishment and full application of the historical paradigm in English dictionary making.

4. The descriptive paradigm of English lexicography

The conceptualization of linguistics underwent radical changes in approaches and dimensions from the late 19th century to the early 20th century, marking a significant transformation in methodology from prescriptivism to descriptivism. The publication of Franz Boas' (1858–1942) *Handbook of American Indian Languages* (1911) marked the germination of descriptive linguistic theories, and their systematic generalization and exposition unfolded with the coming out of

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1704 (Article)

Leonard Bloomfield's (1887–1949) *Language* (1933). Descriptivism, drawing on structural approaches to language, developed out of Bloomfieldian linguistics and on the supposition that description is of greater significance and importance to language pedagogy, research, and training.

Descriptive linguistics, which started to attract serious attention from linguists and language educators in the 1920s, advocates that linguistic description should be based on extensive data, both spoken and written, rather than merely on the written works of the best authors. Linguists should, according to the school of descriptive linguistics, describe the actual use of both spoken and/or written languages and should not prescribe how language should be spoken and written so that a comprehensive, systematic, objective and precise account of the actual use of specific languages over specific periods of time can be provided for certain purposes. All languages and language varieties, whether standard, sub-standard or non-standard, can fulfill communicative functions as long as they are used in speech communities. It is linguists' primary task to make a faithful record of how languages and their varieties are actually used rather than passing judgments upon whether certain uses are right or wrong. Rather than being based on logic and literary superiority, prescriptivism gives prominence to objectivity and systematicity.

Owing to the profound influence of Latin grammar, researches in English grammar and in English dictionary compilation were not able to break through the shackles of prescriptivism, until the rise of Bloomfieldian linguistics caused dramatic changes in the theoretical territories of world linguistics and methodological designing. Against this grand background, Gove's unprecedented masterwork, *Webster's Third*, "a marvelous achievement" and "a monument of scholarship and accuracy", came out in 1961 with brand-new conceptualizations of what English dictionaries should be and what approaches should be adopted to compile such dictionaries. Those concepts eradicated the deep-rooted prescriptive traditions that had been followed for hundreds of years by English dictionary makers and triggered off transformational modifications in the paradigm, notion, and methodology for English dictionary making. The innovations in notions and methodologies were so wide-ranging, so profound and so far-reaching, with no precedents being found in the history of world linguistics and lexicography, that *Webster's Third*, together with its policies of deletions and compiling styles and techniques, met with considerable criticism for its descriptive approach, thus its failure to tell users what proper English was, and its permissiveness. Great controversies were surging among American linguists, lexicographers, dictionary critics and users, and heated debates ensued so that criticisms spurred the creation of *The American Heritage Dictionary of the English Language* (1969), in which usage problems often went to a panel of expert writers for consultation and comments.

This transformation stems chiefly from the following core notions of descriptivism: all language are socially conventionalized systems rather than systems formed through natural laws; the primary step for language research is

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1704 (Article)

observing what really happens when native speakers use the language and making a faithful record of how it is actually used; every language has its unique pronunciation, grammar and vocabulary, which neither logic nor generalized and idealized language can be employed for its description, not even other languages or the diachronic discourse of that language; all languages are dynamic instead of static and are in constant change as long as they are in use by their speakers. Therefore, the so-called "rules" are merely an agreement concerning their current use, and all language use is relative and not absolute. The judgment on whether language use is right or not can only be based on the actual use of language, not on the rules laid down by authorities. These notions, which are the guidelines Gove and his team adopted for compiling their monumental *Webster's Third*, have become the theoretical foundation for the descriptive paradigm of English lexicography. Ever since, the descriptive paradigm has been dominating English dictionary making and research. Descriptivism has played a leading role in the development of English lexicography and has become one of the fundamental principles of modern lexicography.

The seeds of descriptivism are deeply sowed in the minds of present-day linguists and lexicographers. However, the struggle between prescriptivism and descriptivism is far from over. Neither is prescriptivism considered superfluous in language research nor is it ousted from the scene of dictionary making. Just as asserted by Lyons (1968: 43), "it should be stressed that in distinguishing between description and prescription, the linguist is not saying that there is no place for prescriptive studies of language. It is not being denied that there might be valid cultural, social or political reasons for promoting the wide acceptance of some particular language or dialect at the expense of others. In particular, there are obvious administrative and educational advantages in having a relatively unified literary standard." Lexicographers, as well as linguists, have started to assume a serious attitude toward prescriptivism, conduct earnest studies in its application to language pedagogy and dictionary making, and make objective assessments of its role in and influences upon such linguistic and lexicographical activities.

Lexicographers, in particular bilingual lexicographers, are now faced with the challenge of how to implement descriptive ideology in dictionary making while prescriptive traditions are not pulled out of the dictionary-making scene in their entirety. "In 'mainstream' linguistics of recent times scholars have generally claimed that prescription is not a central part of their discipline and even that it is irrelevant to linguistics", but "prescriptive attitudes have far-reaching consequences" (Milroy and Milroy 1985: 5) and have proved to be important and, in some cases, indispensable, such as in language testing and assessment and in dictionary compilation, for two main reasons. First, language is in constant change, with an extraordinarily strong tendency to maintain and regulate its structure, i.e. an instinct of self-maintenance and a process of standardization, which are eventually achieved through language users in response to internal needs for information structuring.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1704 (Article)

"Standardisation is motivated in the first place by various social, political and commercial needs and is promoted in various ways, including the use of the writing system, which is relatively easily standardised; but absolute standardisation of a spoken language is never achieved", and "it seems appropriate to speak more abstractly of standardisation as an ideology, and a standard language as an idea in the mind rather than a reality". "Ultimately, the desideratum is that everyone should use and understand the language in the same way with the minimum of misunderstanding and the maximum of efficiency" (Milroy and Milroy 1985: 22-23). Language standardization is one of the social functions dictionaries should strive to fulfill. That is the most fundamental starting point for launching dictionary projects and also the primary theoretical basis for the effectiveness of dictionaries and their making.

Second, descriptivism has exerted extensive and profound influence upon theoretical inquiries of language, but "the attitudes of linguists ... have little or no effect on the general public, who continue to look to dictionaries, grammars and handbooks as authorities on 'correct' usage" (Milroy and Milroy 1985: 6), for they feel a strong need for rules of "correct" usage and a great necessity of dictionaries and grammars providing such guidance. They are the most convenient authorities to rely on when they are encountered with usage problems and situational perplexities. Any dictionary that excludes judgments about right or wrong usage is doomed to meet with sharp criticism and strong condemnation from its users, especially in the context of cross-cultural communication and foreign language teaching and learning. That has a great deal to do with the deep-rooted tradition of teaching students only standard language in classrooms and testing them only according to norms of standard language.

This tradition is considered necessary and fundamental in the case of foreign language teaching. Non-native learners of a foreign language are generally taught standard foreign languages, and non-standard or informal varieties are strictly excluded from textbooks and classrooms, and therefore, in whatever cases, language learners are denied access to such varieties. Language testing and assessment policies are almost without exception made by the so-called language authorities. Any deviations from the norms or standards prescribed by them are labeled "incorrect" in language testing. The preaching of standard forms of language and the reliance of language learners upon rules of "correct" usage allow for much room for prescriptivism to linger on and to survive in dictionaries, especially in bilingual dictionaries. It can be safely prophesied that it is still a great distance for descriptivism to entirely dominate dictionary making and develop itself from a somewhat idealized model to a dictionary paradigm of fully pragmatic significance.

5. The cognitive paradigm of English lexicography

English lexicography underwent another significant theoretical transformation and shift of focuses in the late 1970s, when Longman broke ground in 1978.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1704 (Article)

Learner's dictionaries, with their origins from the early 20th century and in its wake of Longman, started to mushroom in different forms and in close succession. Their thriving and prosperity pushed dictionary making and research into the era of cognitivism and brought about the perfect integration of dictionary making with language research, cognitive science, language pedagogy, electronic technology, etc.

Learner's dictionaries, also known as pedagogical dictionaries, can be classified in various ways. In the broad sense, they refer to the active-type dictionaries that target all learners of a language for the purpose of linguistic encoding, and in the narrow sense, they refer only to the active-type dictionaries intended for learners of foreign languages or second languages. Learner's dictionaries in the modern sense began to appear as early as the 1930s in the U.K., and the pioneers in this field include Harold Edward Palmer (1877–1949), Michael Philip West (1888–1973), and Albert Sydney Hornby (1898–1978). The early important works, such as *The New Method English Dictionary* (1935) by West and James G. Endicott (1898–1993), *A Grammar of English Words* (1938) by Palmer, *The Idiomatic and Syntactic English Dictionary* (1942) and *The Advanced Learner's Dictionary of Current English* (1948) by Hornby, furnished substantial foundation for the making of English learner's dictionaries and signified the shaping of the dictionary paradigm for the first-generation learner's dictionaries.

The dictionary paradigm for the first-generation learner's dictionaries has its early beginnings in *The New Method English Dictionary*, with a coverage of 23 898 headwords, inclusive of 6 171 phrases and such new words as "crossword" and "vitamin" but exclusive of technical terms and rarely used words. This dictionary is most typically characterized by its defining techniques. All definitions are written as clearly and succinctly as possible, by means of a controlled vocabulary list of 1 490 words and with polysemy explained via synonyms, synonymous expressions and citations. Pictorial illustrations are employed in cases where definitions need to be supplemented and reinforced. Numerous citations are extracted from various data sources to demonstrate the meaning and use of headwords, with due attention given to collocations and fixed usage. It is interesting to note that in the treatment of grammatical information, the compilers focus on its decoding instead of encoding function and provide only meager information items concerning the plural forms of nouns, the comparative and superlative forms of adjectives, the past and past participle forms of verbs, etc., which differs sharply from present-day learner's dictionaries with prominence given to encoding function. However, the whole landscape of grammatical treatment assumed an entirely new look when ALD came out, after continuous supplementation and refinement in Palmer's work and in Hornby's own work, particularly with regard to grammatical rules, verb patterns, grammatical collocation.

The first-generation learner's dictionaries were endowed with brand-new features that made them distinct from previous types of dictionaries. Before them, general monolingual dictionaries were basically intended for native

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1704 (Article)

speakers. With the continuous expansion of the influences of the English language around the world, the special demand for English monolingual dictionaries rose in response to the needs of learners of English as a foreign language. The English teaching experiences West, Palmer and Hornby accumulated over their time overseas, their familiarity with the regularities of foreigners learning English, and their strong awareness of learners' key concerns, major problems and common errors and mistakes in using English provided them with priceless cognitive foundations and designing prerequisites for English learners' dictionary making. The theories of controlled defining vocabulary, phraseology, and pedagogical grammar, which drew serious attention from lexicographers and language educators, became the theoretical source and energy for the emergence and development of English learner's dictionaries in the early stage.

The emergence of the second-generation learner's dictionaries were marked by the publication of the second (1963) and third (1974) editions of ALD and the first edition of Longman in 1978, which ushered in a new era of dictionary making being geared to the special needs of linguistic output. By the 1980s, the third-generation learner's dictionaries were ready to make their appearance, as a result of the rapid development of modern linguistics (especially pragmatics, cognitive linguistics, applied linguistics, corpus linguistics, computational linguistics, etc.), the more in-depth studies in dictionary use and user cognition and the timely introduction of mature electronic information technology into dictionary making. The third generation was signified and represented by *The Oxford Advanced Learner's Dictionary of Current English* (Fourth Edition, 1989), *Longman* (Second Edition, 1987) and John Sinclair's *Collins Cobuild Dictionary of the English Language* (1987).

By the 1990s, English learner's dictionaries entered into an epoch of thriving and prosperity. New editions of OALD, Longman and CCD emerged one after another, and new dictionary brands, such as Cambridge and Chambers, started to squeeze into the learner's dictionary market, owing to the revolutionary tides already surging in the lexicographical circles and the irresistible temptation of high profitability and enormous market potentials. The year 1995 is of special significance in the history of English learners' dictionary making in that it witnessed the almost simultaneous coming-out of "The Big Four", i.e. the first edition of *Cambridge International English Dictionary*, as well as new editions or new reprints of *OALD*, *Longman* and *CCD*. Beyond the Atlantic Ocean, another "Big Four" also made their appearance on the American dictionary market, i.e. *The American Heritage Dictionary* (College Edition, 1982), *The Random House College Dictionary* (1966–1975), *Webster's New Collegiate Dictionary* (1949–1976) and *Webster's New World Dictionary of American English* (1972). English learner's dictionaries, along with the collegiate dictionaries in America, combined to forge an era of userism and the cognitive paradigm for English dictionary making.

The cognitive paradigm of English lexicography is a natural outcome of integrated developments in theorization of cognitive science, cognitive linguistics, lexicography and foreign language pedagogy. Cognitive science, which examines human mind and the way it works and analyzes the nature, tasks and functions of cognition, can trace its origins to the studies in the nature of human knowledge and the observation and thinking of human mind conducted by ancient Greek philosophers, such as Plato (427 B.C.–347 B.C.) and Aristotle (384 B.C.–322 B.C.), and to the findings in human mind explorations by René Descartes (1590–1650), Baruch de Spinoza (1632–1677) and other philosophers. It is an interdisciplinary field that incorporates accomplishments in philosophy, psychology, neuroscience, sociology, anthropology, biology, computer science, artificial intelligence, linguistics and other related disciplines, as a complete and sound understanding of human mind and its interactions with the surrounding world can only be obtained from a combination of diverse dimensions.

The cognitive dictionary paradigm is based on such cognitive linguistic notions: language is not a self-contained vacuum system, linguistic competence is part of human cognitive capabilities, and language description must draw inferences from cognitive processes; linguistic structure has something to do with the conceptual structure, knowledge structure, discourse function and practical experience of the humanity and uses them as motivation to frame; syntax is not a self-perfecting system and intertwined with vocabulary and semantics, vocabulary, morphology and syntax are continuums constituting a semiotic body; semantics is not merely objective truth conditions but is closed associated with the subjective mind and the infinite knowledge system of humankind; dictionary making and use are socio-cultural activities that highlight the natural process of linguistic cognition and the mental representation of vocabulary acquisition.

In practice, the cognitive paradigm of English lexicography starts from the links and processes of users' linguistic cognition. It adopts cognitive approaches and examines such dimensions as formal structure, categorical structure, valence structure and distributional structure to expound headwords in the dictionary and how they are acquired by users. It attempts to decipher the flow-process diagram of cognition, explore the lexical mental representation of potential dictionary users, the cognitive process of dictionary consultation, the needs and skills of information look-up, the learning strategies of dictionary use, etc. so as to enhance the efficiency of lexical acquisition. All this entails the shift in dictionary making from compiler-centered to user-centered, from decoding-focused to encoding-focused, and from consultative look-up to productive association. English learner's dictionaries, after undergoing three generations of development, have become relatively mature and at the cutting edge of the theory and practice of world learner's lexicography.

English learner's lexicography started to sublate, from the time of its formation, the prescriptive and diachronic approaches of traditional linguistics and turned to the modern synchronic descriptive approach. It borrows ideologies from various fields, including the structural behavioral theory, the trans-

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1704 (Article)

formative generative theory, the theory of cognition and communication in pragmatics, the theory of second language acquisition, and even the fashionable theories of prototypic categories and metaphor in cognitive linguistics, which are reflected to some extent in the compiling strategies of *Macmillan English Dictionary for Advanced Learners* (2002). It strictly restricts the defining vocabulary limit and the density of lexical coverage, augments the number of grammatical information items and citations that facilitate linguistic production, gives primary prominence to supplementary functions that are conducive to both encoding and decoding, including language notes, usage guides, guidewords, signposts and so on, and strengthens the role of electronic information technology and corpuses in selecting headwords, senses, controlled defining vocabulary, citations, usage explanation, and variety indication, and in revision, augmentation and supplementation. All this highlights the conspicuous characteristics of English learner's dictionaries — cutting edge, flexible, handy, easy to use, efficiency-focused and user-oriented.

WordNet is an online dictionary, a large and extraordinary lexical database of English that bears striking resemblance to a thesaurus, with nouns, verbs, adjectives and adverbs being grouped into 117 000 sets of cognitive synonyms (synsets) on the semantic basis. It is produced by a team of linguists, lexicographers, psychologists and computer engineers in Princeton University and is available at www.wordnet.princeton.edu. It employs cognitive principles in its making to such an extent that each synset expresses "a distinct concept" and "is linked to other synsets by means of a small number of 'conceptual relations', "contains a brief definition ('gloss') and, in most cases, one or more short sentences illustrating the use of the synset members". "The main relation among words in *WordNet* is synonymy". "The resulting network of meaningfully related words and concepts can be navigated with the browser" (see *WordNet* website).

WordNet features theoretical breakthroughs, independent compilation through research and cognitive representation of lexical consultation and acquisition. It is a perfect integration of traditional techniques of treating lexicographical information, modern online information processing technology and research findings in psychology and cognitive linguistics. Its most conspicuous innovation resides in its organization of lexical information, linguistic knowledge and the whole text according to conceptual relations, sense relations and in some cases even senses proper, rather than word forms to simulate and reflect mentally human cognition of lexical items. "*WordNet*'s structure makes it a useful tool for computational linguistics and natural language processing" (see *WordNet* website). *WordNet* has proved to be of rich theoretical implication and huge practical value to studies in computation linguistics, psycholinguistics, cognitive linguistics, mental representation of lexical acquisition, lexical teaching and learning, online database building, online lexicographical information arrangement and presentation, analysis of automatically generated text, application of artificial intelligence, and natural language processing.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1704 (Article)

6. Conclusion

After over 1 500 years of evolution, English has become a truly globalized language. Owing to its rapid expansion into the international community in the past centuries and the strengthened status in the international arena today, English dictionaries have consolidated their ever-increasing influences upon both theory and practice of world lexicography. In less than 500 years, English dictionaries have completed their evolution from their archetype to prescriptivism, historicism, descriptivism and then to cognitivism, which amply demonstrate the sociocultural and interdisciplinary nature of dictionary making and research, the interactive relations between language and dictionary, dictionary and culture, dictionary making and user needs, dictionary design and user research, and finally dictionary use and language pedagogy. All these combine to present the evolutional chain of English dictionary paradigms, a complete, coherent and unified portrayal of the trace English dictionaries follow in their development up to the present times.

Major References

Algeo, John and Thomas Pyles. 2009. *The Origins and Development of the English Language.* Belmont, CA.: Wadsworth.

Atkins, Sue and Michael Rundell. 2008. *The Oxford Guide to Practical Lexicography.* Oxford: Oxford University Press.

Bartsch, Renate. 1987. *Norms of Language. Theoretical and Practical Aspects.* London/New York: Longman.

Béjoint, Henri. 2010. *The Lexicography of English. From Origins to Present.* Oxford: Oxford University Press.

Collison, Robert L. 1982. *A History of Foreign-Language Dictionaries.* London: Andre Deutsch.

Granger, Sylviane and Magali Paquot (Eds.). 2012. *Electronic Lexicography.* Oxford: Oxford University Press.

Green, Jonathon. 1996. *Chasing the Sun. Dictionary-Makers and the Dictionaries They Made.* London: Jonathan Cape.

Hudson, Richard. 1988. The Linguistic Foundations for Lexical Research and Dictionary Design. *International Journal of Lexicography* 1(4): 287-312.

Hüllen, Werner. 2006. *English Dictionaries 800–1700: The Topical Tradition.* Oxford: Clarendon.

Johnson, Samuel. 1747. *The Plan of an English Dictionary.* Edited by Jack Lynch [online]. Available at: https://jacklynch.net/Texts/plan.html.

Johnson, S. 1755. *Johnson, Preface to the Dictionary. From Samuel Johnson,* A Dictionary of the English Language *(London, 1755).* Edited by Jack Lynch [online]. Available at: https://jacklynch.net/Texts/preface.html.

Kaster, Robert. 2009. Latin Lexicography. *The Classical Review* 59(1): 169-171.

Krache, Robert. 1975. *The Story of the Dictionary.* New York: Harcourt Brace Jovanovich.

Kuhn, Thomas. 1970. *The Structure of Scientific Revolutions.* Chicago: University of Chicago Press.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)
http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1704 (Article)

Landau, Sidney I. 1989. *Dictionaries*: *The Art and Craft of Lexicography.* Cambridge: Cambridge University Press.

Liberman, Anatoly. 2009. English Etymological Dictionaries. Cowie, A.P. (Ed.). 2009. *The Oxford History of English Lexicography. Volume II:* 269-289.Oxford: Oxford University Press.

Lyons, John. 1968. *Introduction to Theoretical Linguistics*, Cambridge: Cambridge University Press.

McArthur, Tom. 1986. *Worlds of References-lexicography, Learning and Language: From Clay Tablets to Computer.* Cambridge: Cambridge University Press.

Milroy, James. 1992. *Linguistic Variation and Change.* Oxford: Blackwell.

Milroy, James and Lesley Milroy. 1985. *Authority in Language —Investigating Language Prescription and Standardization.* Routledge and Kegan Paul.

Morton, Herbert C. 1994. *The Story of Webster's Third: Philip Gove's Controversial Dictionary and Its Critics.* Cambridge/New York: Cambridge University Press.

Mugglestone, Lynda. 2005. *Lost for Words: The Hidden History of the Oxford English Dictionary.* New Haven, CT/London: Yale University Press.

Murray, James A.H. 1900. *The Evolution of English Lexicography. (The Romanes Lecture 1900.)* London: Henry Frowde / Oxford: Clarendon Press.

Ogilvie, Sarah. 2020. *The Cambridge Companion to English Dictionaries*. Cambridge: Cambridge University Press.

Prinsloo, D.J. 2005. Electronic Dictionaries Viewed from South Africa. *Hermes, Journal of Linguistics* 34: 11-35.

Reddick, Allen. 1990/1996. *The Making of Johnson's Dictionary 1746–1773.* Cambridge: Cambridge University Press.

Rosch, E. 1975. Cognitive Representations of Semantic Categories. *Journal of Experimental Psychology: General* 104(3): 192-233.

Sauer, Hans. 2009. Glosses, Glossaries, and Dictionaries in the Medieval Period. Cowie, A.P. (Ed.). 2009. *The Oxford History of English Lexicography. Volume I: General-purpose Dictionaries:* 17-40. Oxford: Oxford University Press.

Starnes, DeWitt T. and Gertrude E. Noyes. 1991. *The English Dictionary from Cawdrey to Johnson 1604–1755.* [Second edition with an introduction and new bibliography by Gabriele Stein.] Amsterdam/Philadelphia: John Benjamins.

Stathi, E. 2006. Latin Lexicography. Brown, K. (Ed.). 2006. *Encyclopedia of Language and Linguistics. Volume 6:* 723-724. Second edition. Oxford: Elsevier.

Travis, C. 2000. *Unshadowed Thought: Representation in Thought and Language.* Cambridge, MA: Harvard University Press.

Trench, Richard. 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.

Wells, Ronald A. 1973. *Dictionaries and the Authoritarian Tradition: A Study in English Usage and Lexicography.* The Hague: Mouton.

Yong, Heming. 2022. *English Lexicography from British Tradition to World Englishes.* New York: Peter Lang.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (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.

http://lexikos.journals.ac.za; https://doi.org/10.5788/32-2-1705 (Volle uitgawe / Full volume)

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**.